

**Louise-Amélie Cougnon
and Cédrick Fairon (eds.)**

SMS Communication

BENJAMINS CURRENT TOPICS 61

SMS Communication

Benjamins Current Topics

Special issues of established journals tend to circulate within the orbit of the subscribers of those journals. For the Benjamins Current Topics series a number of special issues of various journals have been selected containing salient topics of research with the aim of finding new audiences for topically interesting material, bringing such material to a wider readership in book format.

For an overview of all books published in this series, please see
<http://benjamins.com/catalog/bct>

Volume 61

SMS Communication. A linguistic approach

Edited by Louise-Amélie Cougnon and Cédric Fairon

These materials were previously published in *Linguisticae Investigationes* 35:2 (2012)

SMS Communication

A linguistic approach

Edited by

Louise-Amélie Cougnon

Cédrick Fairon

Catholic University of Louvain

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

SMS communication : a linguistic approach / edited by Louise-Amélie Cougnon, Cédric Fairon.

p. cm. (Benjamins Current Topics, ISSN 1874-0081 ; v. 61)

“These materials were previously published in *Linguisticae Investigationes* 35:2 (2012).”

Includes bibliographical references and index.

1. Communication--Technological innovations. 2. Communication and culture--Technological innovations. 3. Language and languages--Usage. 4. Interpersonal communication--Technological innovations--Social aspects. 5. Cell phones--Social aspects. I. Cougnon, Louise-Amélie. II. Fairon, Cédric. III. Title: Short message service communication.

P96.T42S74 2014

302.23'445--dc23

2014004754

ISBN 978 90 272 0280 2 (Hb ; alk. paper)

ISBN 978 90 272 7030 6 (Eb)

© 2014 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Acknowledgements	VII
Foreword	1
<i>David Crystal</i>	
Introduction	3
<i>Louise-Amélie Cougnon and Cédric Fairon</i>	
Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques	11
<i>Pierre Accorsi, Namrata Patel, Cédric Lopez, Rachel Panckhurst and Mathieu Roche</i>	
SMS experience and textisms in young adolescents: Presentation of a longitudinally collected corpus	29
<i>Josie Bernicot, Olga Volckaert-Legrier, Antonine Goumi and Alain Bert-Erboul</i>	
Automatic or Controlled Writing? The Effect of a Dual Task on SMS Writing in Novice and Expert Adolescents	47
<i>Céline Combes, Olga Volckaert-Legrier and Pierre Largy</i>	
Development of SMS language from 2000 to 2010: A comparison of two corpora	67
<i>Úrsula Kirsten Torrado</i>	
Texto4Science: A Quebec French database of annotated text messages	87
<i>Philippe Langlais and Patrick Drouin</i>	
SMS communication as plurilingual communication: Hybrid language use as a challenge for classical code-switching categories	111
<i>Etienne Morel, Claudia Bucher, Simona Pekarek Doehler and Beat Siebenhaar</i>	
French text messages: From SMS data collection to preliminary analysis	141
<i>Rachel Panckhurst and Claudine Moïse</i>	

A sociolinguistic analysis of transnational SMS practices: Non-elite multilingualism, grassroots literacy and social agency among migrant populations in Barcelona	169
<i>Maria Sabaté i Dalmau</i>	
Negation marking in French text messages	191
<i>Elisabeth Stark</i>	
“i didn’t spel that wrong did i. Oops”: Analysis and normalisation of SMS spelling variation	217
<i>Caroline Tagg, Alistair Baron and Paul Rayson</i>	
<i>Lol, mdr</i> and <i>ptdr</i> : An inclusive and gradual approach to discourse markers	239
<i>Deniz Uygur-Distexhe</i>	
Index	265

Acknowledgements

We would like to thank Bernadette Dehottay and Jennifer Thewissen for their help in the editing process of this volume and the *Institut Langage et Communication* for their support of this publication.

Scientific Committee

Georges Antoniadis	Laboratoire de linguistique et didactique des langues étrangères et maternelles, Université Stendhal de Grenoble, France
Josie Bernicot	Centre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers, France
Noémi Clothilde Boubel	Centre de traitement automatique du Langage, UCL, Belgium
Jean-Léon Bouraoui	Prometil, Toulouse, France
Louise-Amélie Cougnon	Centre de traitement automatique du Langage, UCL, Belgium
Orphée De Clercq	Language and translation technology team, Hogeschool Gent, Belgium
Marie-Catherine de Marneffe	Department of Linguistics, Stanford University, United States
Thierry De Smedt	École de communication, UCL, Belgium
Anne Dister	Facultés universitaires Saint-Louis & UCL, Belgium
Patrick Drouin	Oservatoire de Linguistique Sens-Texte, Université de Montréal, Canada
Cédrick Fairon	Centre de traitement automatique du Langage, UCL, Belgium
Thomas François	Institute for Research in Cognitive Science, University of Pennsylvania, United States
Tita Kyriakopoulou	Laboratoire d'informatique Gaspard-Monge, Université Paris-Est Marne-la-Vallée, France
Laurent Kevers	Centre de traitement automatique du Langage, UCL, Belgium
Jean Klein	UCL, Belgium
Gudrun Ledegen	Département Communication, Université Rennes 2, France
Mario Marcon	University of Udine, Department of Foreign Languages and Literatures, Italy
Elizabeth Marshman	School of Translation and Interpretation, University of Ottawa, Canada
Carla Meskill	Department of Educational Theory and Practice, University at Albany, United States
Takuya Nakamura	Université Paris-Est Marne-la-Vallée, France

Rachel Panckhurst	Praxiling UMR 5267 CNRS — Université Paul-Valéry — Montpellier 3, France
Simona Pekarek Doehler	Institut des sciences du langage et de la communication, Université de Neuchâtel, Switzerland
Mathieu Roche	Laboratoire d’Informatique, de Robotique et de Microélectronique, Université Montpellier 2, France
Elisabeth Stark	Romanisches Seminar, Universität Zürich, Switzerland
Caroline Tagg	Centre for English Language Studies, University of Birmingham, Great-Britain
Jennifer Thewissen	Centre for English Corpus Linguistics, UCL, Belgium
Stéphanie Weiser	Centre de traitement automatique du Langage, UCL, Belgium
Anastasia Yannacopoulou	Technological Institute of Crete, Hellenic Open University, Greece, Ergonotics, France
François Yvon	Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur, Université Paris Sud, France

Foreword

David Crystal

SMS is a genre in which the ideational, identifying, and ludic functions of language combine in a climate of rapid technological change. As such, it presents linguists with a special challenge. One begins a research project, confident that one is at the cutting-edge of language study, and concludes it a couple of years later only to find that actually what one has been doing is historical linguistics! Those who began to study Twitter between 2006 and 2009 know exactly what I mean. When Twitter changed its prompt (from 'What are you doing?' to 'What's happening?'), the change of orientation from an inward-looking to an outward-looking perspective resulted in a significant shift both in linguistic content and linguistic form. Messages that previously focussed solely on 'I', 'me', and 'we' were now presenting narratives that prioritized 'it', 'he', and 'she'. Generalizations about Twitterspeak in its first three years no longer applied to its next three. And who knows what will happen in the three years after that?

Linguists therefore have to get their act together, otherwise they will miss the opportunity to analyse the properties of a new medium of communication in its earliest stages of evolution, and this is why the present volume is timely. There are no precedents here: linguists weren't around when speech, writing, and signing developed. But here is electronically mediated communication, in its various technological outputs, displaying a veritable explosion of linguistic innovation, and offering exciting opportunities for description and analysis. It is a field where it is virtually impossible to do anything other than original research. With some fields (Shakespeare comes to mind), investigators can struggle to find a new angle, simply because of the amount of study that has already been carried out. That is not the case here. Every language, dialect, and social group develops its own brand of electronic communication, and very few of these varieties have had their linguistic character analysed — or even recorded.

Several of the papers in this volume acknowledge this lack of research. SMS has suffered especially from the mythology which surrounded its arrival in the early 2000s. Linguists who were concerned to establish the properties of this new genre found that much of their energy was channelled into corrective presentations.

I have lost track of the number of hours I spent trying to counter the bizarre characterizations of SMS that appeared in the media, or the hysterical harangues that emanated from radio and newspaper journalists. I would far rather have devoted that energy to the kind of research that we see presented in these pages. Specifically linguistic initiatives were also sidelined by the way popular debate focused on the imagined consequences of SMS for children in school, especially in relation to spelling. It proved far easier to get a grant to explore educational outcomes than to provide linguistic descriptions.

We need these descriptions, and their associated analyses, and nowhere more urgently than in relation to SMS, for this is a genre whose future is unclear. Although the increase in mobile communication is one of the major trends at present, especially in parts of the world where Internet access via conventional terminals is limited by poor wired connections, we are also encountering another trend, in which the proportion of oral/aural to graphic electronic communication is steadily rising, and a world where speech-to-text software is high quality and routine is just around the corner. What will happen to SMS then?

Looking back at the papers in this volume in ten year's time, I think they will be of value because they present a synchronic snapshot of the kinds of things that were happening during the early days of a new genre. Some of the features they describe are likely to be transient, in the sense that they will either be replaced by features reflecting new technology or be dropped because they are no longer of interest to users. I remember going into secondary schools in Britain in the early 2000s and finding that all the children were experimenting with what they felt to be exciting nonstandard abbreviatory conventions. Continued visits through the decade and into 2012 saw many changes. In a recent visit to one school, the students had collected text messages from each grade, and found that, the younger the students, the more they used abbreviated forms. The abbreviations had largely disappeared from older children's texts. 'We used to do that when we were younger', said one 17-year-old. 'Textisms are naff [i.e. unfashionable]', said another. 'I stopped using them when my parents started to', said a third.

A decade is a short time in linguistics.

Introduction

Louise-Amélie Cougnon and Cédrick Fairon

Université catholique de Louvain, Institut Langage et Communication,
Centre de traitement automatique du langage

The worldwide¹ adoption of Computer-Mediated Communication (CMC) in the last decades has led to a series of questions, including for instance the impact of new practices on social behaviour, spelling performance and language development. The scientific world started concentrating on the social and psychological effects of such practices as early as the 1970s (e.g. M. Turoff and S.R. Hiltz 1977). The results showed a double and paradoxical trend: increasing extrovert behaviour (people increasingly communicating with each other) and a serious withdrawal movement (people cutting themselves off from the non-computer world, cyber-dependency...) (F. Bamba and S.J. Barnes 2007, L. Leung 2007, R. Wei *et al.* 2010, B. Moynihan *et al.* 2010).

As we have just celebrated the 20th anniversary of SMS communication (1992–2012), the social and technological contexts of use have noticeably evolved over years. Early studies have shown that the characteristics of the tools used for scriptural electronic communication may have a direct influence on the text produced as they impose various constraints on the user: some tools produce messages of a limited size only (SMS) whereas other impose almost no limitation (emails); some are used with a full-size computer keyboard (emails, chat, forum) and others with a keypad (SMS); some are designed for instant communication (chat) and others have a longer delivery process sometime involving review and moderation (forum); some are dedicated to exchanges between two people and other are specifically for large scale communication (forum, blog, etc.); some require that you sit in front of a computer and some others are available on your cell phone. These technical and situational aspects represent facets that researchers have used to distinguish SMS texting from chats, emails and other forms of CMC texts (Herring

1. This development is not socially nor geographically homogeneous. As Thurlow (2003, 1) points out: « which is not to say that this technology is properly global; worldwide patterns of [media development] necessarily follow the socioeconomic contours of which distinguish the 'media rich' and 'media poor' generally ».

2007). There is no doubt that texting practices vary across these communication tools and that what is considered as acceptable in one type of media is not always acceptable in another. To verify this claim, one just needs to type in a phrase such as “no SMS/text language” in a web search engine. This query will return thousands of utterances stating “no SMS/text language in emails, reviews, posts, etc.,” mostly appearing on forum and blogs that include rules about texting in their “netiquette”.

Today, a global phenomenon called “digital convergence” is reshaping the world of information technologies, communications, electronics and entertainment. This unstoppable trend goes in the direction of a general merging of devices and services that were formerly independent such as phones, television, the radio, cameras and computers. Services are interconnected and available on various tools. Smartphones clearly illustrate this: they can either give access to the Internet through a cellular network or connect to a wireless hotspot to give access to a voice over ip provider on the Internet. Moreover, the picture you take with your phone can be uploaded on your blog or sent by email regardless of the way you connected to the Internet (using a cellular or wireless connection). These changes also affect CMC tools. With the spreading of the Smartphone market, they are becoming more mobile and ubiquitous.

In this context, it is likely that “digital convergence” will lead to a convergence of scriptural practices in CMC. Visible traces of this evolution are already noticeable. For instance, the automatic signature added by email software on smartphones (“Sent from my iPhone/BB/phone”) sounds like a warning. The user is informed that the email might not respect the general convention of the text genre, because it was sent from a specific device (with a small keypad) or in a situation that for instance required to be prompt, which can result in a more succinct, and direct text or one which contains abbreviations or typos. The Facebook messaging system is another emblematic example of the ongoing mutation of CMC tools. It integrates in a unified system soberly called “Messages” texts that can be sent or received by email, SMS or through the Facebook instant messaging system. All these messages are combined in one feed that shows how much the communication tools are integrated. A discussion can start by chat, be pursued by SMS and be concluded by an email, for example. It is simple and it meets the expectations of a young public which find traditional email far too formal.²

Unfortunately, *linguistic* issues pertaining to computer-mediated communication have, to date, been generally underresearched. Some studies which have dealt with the linguistic aspects include R. Panckhurst (1997), S. Herring (1998), J. Anis (1999), L. Mondada (1999) and J. Gains (1999) and were dedicated to email and

2. A ComScore study (2010) showed that in the USA, email was drastically abandoned (-59%) by adolescent who preferred SMS communication and social networks.

forum writing mainly. Questions of style, genre and register were at the center of these studies.

Short Message Service (SMS) is, along with instant messaging, blogging, emailing, social networking, and virtual worlds, one of the modern forms of worldwide communication.

Because mobile phones and mobile subscriptions (mostly SMS packs) have become increasingly cheaper, “texting” (D. Crystal 2008) has emerged as the most widely used form of CMC.³ Despite this exponential social phenomenon, linguistic studies based on text messages corpora appeared quite late (*i.e.* towards the late 1990s and 2000s) and, at first, were mainly dedicated to the study of SMS abbreviation techniques (C. Fairon *et al.* 2006; C. Frehner 2008; R. Panckhurst 2009). Thanks to this pioneering research, we are now aware that linguistic phenomena tightly associated with SMSs include phonetic and numeral scripts, abbreviations and capital letters, and the use of new code systems such as emoticons. However, we still know little about SMS lexical and dialogical properties, as well as syntactic trends; knowledge about the impact of SMS practices on more formal writing practices (dictation, professional reports...) is scarcer still.

Some studies about SMSs initiated a move towards more varied research subjects: for example, in the field of interaction (J. Walkowska 2009; D. Laursen 2005) or in that of lexicology (C. Fairon *et al.* 2006; L.-A. Cougnon and C. Fairon 2010, C. Tagg 2009); syntactic results have also been highlighted by E. Stark (2011). Some sociolinguistic trends have been pointed out by K. Vold Lexander (2007), L.-A. Cougnon (2012), E.-L. Kasesniemi and P. Rautiainen (2001), D. Laursen (2005), J. Bernicot *et al.* (2012) and C. Thurlow (2003) among others. The subject of SMSs also caught the attention of the Natural Language Processing field which focused on aspects such as word prediction, spelling normalization, or speech synthesis (Y. How and M.-Y. Kan 2005; M. Schneider-Hufschmidt 2005; F. Yvon 2008; R. Beaufort *et al.* 2008; E. Guimier De Neef and S. Fessard 2007; C. Kobus *et al.* 2008, R. Beaufort *et al.* 2010; J. Oliva *et al.* 2013). All of these studies are pioneering and there is a need for new results that will confirm their results and which can be applied generally.

The tendency to shy away from the linguistic analysis of text messages can partly be explained by the difficulty involved in collecting SMS corpora. Such a collection is made difficult by privacy matters (compared to blogs and website correspondence for instance) but also by more technical aspects (data transmission from mobile phones to databases). These restrictions have led to the compilation of

3. An estimated 8 trillion text messages have been sent in 2011. Although, in the last few years, SMS communication came up against fierce competition from free messaging apps on computers and smartphones, analysts say it will remain a dominant tool on mobiles.

a large number of corpora which are limited in size (several hundred text messages) or properties (not written in real-life situations, written for specific research purposes,, written by the researcher's relatives, manually transcribed...). Since the 2000s, however, some projects have started collecting large and real-life SMS corpora.

An international list of SMS corpora has been drawn up in T. Chen and M.-Y. Kan (2011): it includes the 85,870-message corpus of W. Liu and T. Wang, 2010 and the 10,000-SMS corpus of C. Tagg (2009). The international sms4science project⁴ is the largest real-life SMS corpus to date: it consists of more than 220,000 SMSs. Four countries have already taken part in the project: France (La Reunion: 12,000 SMSs; Montpellier: 90,000 SMSs; Grenoble 4,500 SMSs), Canada (20,000 SMSs), Switzerland (26,000 SMSe) and Belgium (70,000 SMSs). Each local coordinator

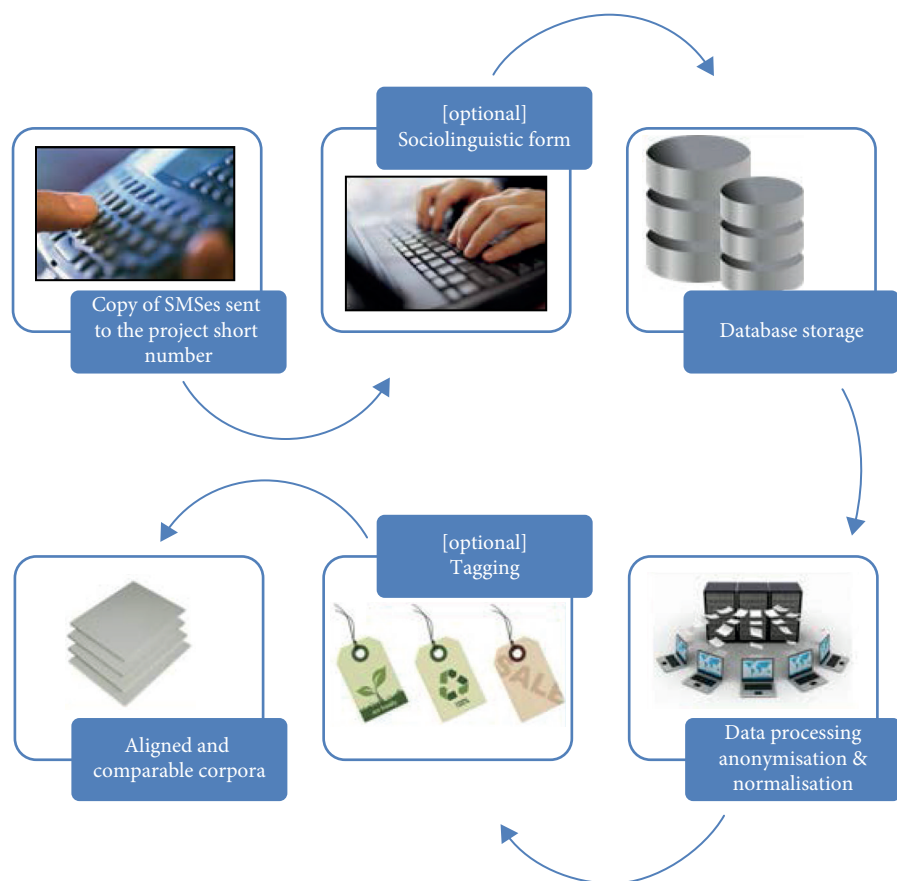


Figure 1. The sms4science project: methodological steps

4. Readers are referred to www.sms4science.org for a detailed bibliography of existing studies about SMSs. New references are welcome and can be sent to louise-amelie.cougnon@uclouvain.be.

gathered the regional data following a strict collection methodology in order to obtain comparable SMS corpora. SMS users were asked to send a copy of the text messages that they had sent to a real addressee in a genuine communication situation. They could do this via the sms4science project short number. The methodology is detailed in Figure 1.

The sms4science methodology also enabled sociolinguistic information about SMS users to be gathered: e.g. age, sex, mother tongue, number of other languages spoken and SMS-specific habits (number of SMSs sent per week, use of the predictive text entry...). At present, the data include approximately 4,000 sociolinguistic SMS-user profiles linked to their respective text messages.

Such corpus collections enable a wider panel of studies to be carried out. To this day, SMS-oriented work with a linguistic focus is sorely missing. The volume edited by R. Harper *et al.* (2005) was among the first to delve into this issue, but included studies based either on small amounts of data or non-linguistic research questions.

The aim of this volume is threefold: to present recent linguistic research in the field of SMS communication; to inform the reader about existing large SMS corpora and processing tools and, finally, to show the many linguistic aspects that can be studied via a corpus of SMSs. This volume was first published as a special issue of *Linguisticae Investigationes* and appears in this book in a reviewed version.

The papers are presented in alphabetical order following the first author's surname. The volume offers two papers related to methodological matters: *Panckhurst and Moïse* retrace the organization of their data collection, the elaboration of their sociolinguistic questionnaire, text message data processing procedures, and present a future possible methodology for conversational SMS extraction; in the second paper, *Langlais and Drouin* share their experience of the annotation of large corpora, both at the automatic and linguistic levels.

The volume also includes sociolinguistic perspectives: *Bernicot et al.* analyse the concept of variation as testified in the density of textisms according to age and sex; *Sabaté* focuses on the linguistic features of transnational SMS talk in a group of migrants based in Barcelona. *Combes et al.* carry out a psycholinguistic analysis to show that spelling in SMS contexts differs according to SMSs expertise and participant attention.

Kirsten adopts a diachronic perspective and details the development of SMS language during the last decade. The author concludes that shortened versions of words are not always the aim of respelling in SMS contexts.

Syntactic aspects are broached by *Stark*, whose paper investigates the question of negation marking in SMS contexts. *Uygur-Distexhe* studies the function of the three initialisms, viz. *lol*, *mdr* and *ptdr*, that work syntactically and semantically as discourse markers. *Morel et al.* underline the need for an adaptation of the classical

methodological apparatus of code-switching studies in SMS communication. In addition to switching between natural languages, code-switching must also be understood to include switching between natural languages and language varieties, switching between styles and switching between languages and other semiotic systems.

Finally, a natural language processing part includes a paper by *Tagg et al.* who present the contributions of a qualitative analysis to the automatic normalization of spelling in SMSs. The second paper in this category is by *Accorsi et al.* who describe *Seek&Hide*, a text message processing tool that automatically performs the anonymisation of a corpus.

References

- Anis, Jacques (dir.). 1999. *Internet, communication et langue française*, Paris, Hermes Sciences Publications.
- Bamba, Fatim; Stuart J. Barnes. 2007. Sms advertising, permission and the consumer: a study. *Business Process Management Journal* 13(6): 815–829
- Beaufort, Richard; Sophie Roekhaut; Cédric Fairon. 2008. Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition ». *Proceedings of JADT 2008*: 155–166.
- Beaufort, Richard; Sophie Roekhaut; Louise-Amélie Cougnon; Cédric Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*: 770–779, Uppsala. Association for Computational Linguistics.
- Bernicot, Josie; Olga Volckaert-Legrier; Goumi, A. ; Alain Bert-Erboul. 2012. Forms and functions of SMS messages: A study of variations in a corpus written by adolescents. *Journal of Pragmatics* 44: 1701–1715.
- Chen, Tao; Min-Yen Kan. 2011. Creating a live, public short message service corpus: The nus sms corpus. *Computing Research Repository* abs/1112.2468.
- Cougnon, Louise-Amélie; Cédric Fairon. 2010. La néologie dans l'écrit spontané. Etude d'un corpus de SMS en Belgique francophone". *Actes du Congrès International de la néologie dans les langues romanes*. Barcelone. Série Activitats 22: 1139–1154.
- Cougnon, Louise-Amélie. 2012. *L'écrit sms. Variations lexicale et syntaxique*. PhD thesis defended on September 15th 2012. Dir. : Cédric Fairon. Université catholique de Louvain. Belgium
- ComScore. 2011. The 2010 U.S. Digital Year in Review, A recap of the year in digital media. http://www.comscore.com/Press_Events/Presentations_Whitepapers/2011/2010_US_Digital_Year_in_Review
- Crystal, David. 2008. *Txtng: the Gr8 Db8*. Oxford University Press.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier. 2006. *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête Faites don de vos SMS à la science*, Les Cahiers du Cental, Presses universitaires de Louvain.
- Frehner, Carmen. 2008. *Email — SMS — MMS : The linguistic creativity of asynchronous discourse in the new media age*. Peter Lang.

- Gains, Jonathan. 1999. Electronic Mail—A New Style of Communication or Just a New Medium? An Investigation into the Text Features of E-Mail. *English for Specific Purposes* 18(1): 81–101.
- Guimier De Neef, Emilie; Sébastien Fessard. 2007. Évaluation d'un système de transcription de SMS. In *Acte du 26e Colloque international Lexique Grammaire*, Bonifacio, 2–6 octobre: 217–224.
- Harper, Richard; Leysia Ann Palen and Alex Taylor (eds). 2005. *The inside text: Social, cultural and design perspectives on SMS*. Springer.
- Herring, Susann. 1998. Le style du courrier électronique : variabilité et changement. *Revue d'aménagement linguistique* 84–85: 9–16.
- Herring, Susann. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet*. <http://www.languageatinternet.de/articles/2007/761>
- How, Yijue; Min-Yen Kan, M. Y. 2005. Optimizing predictive text entry for short message on Mobile Phones. *Human Computer Interfaces International (HCII 05)*. Las Vegas, USA. <http://faculty.washington.edu/thurlow/papers/Thurlow%282003%29-DAOL.pdf>.
- Kasesniemi, Eija-Liisa; Pirjo Rautiainen. 2001. Life in 160 characters: The text message culture of Finnish teenagers. In Katz, J. E.; Aakhus, M. (eds). *Perpetual contact : Mobile communication, private talk, and public performance*: 170–192. Cambridge University Press.
- Kobus, Catherine; François Yvon ; Géraldine Damnat. 2008. Normalizing SMS: are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics* 1: 441–448, Manchester. Association for Computational Linguistics.
- Laursen, Ditte. 2005. Please reply! The replying norm in adolescent SMS communication. In Harper, R., Palen, L.A.; Taylor, A.S. (eds). *The Inside Text*: 53–73. Springer, Dordrecht, kluwer aca edition.
- Leung, Louis. 2007. Unwillingness-to-communicate and college students' motives in SMS mobile messaging. *Telematics and Informatics* 24(2): 115–129.
- Liu, Wuying; Ting Wang. 2010. Index-based Online Text Classification for SMS Spam Filtering. *Journal of computers* 5(6): 844–851.
- Mondada, Lorenza. 1999. Formes de séquentialité dans les courriels et les forums de discussion. Une approche conversationnelle de l'interaction sur Internet. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication* 2(1): 3–25.
- Moynihan, Brynn; Sertan Kabadayi; Mark Kaiser. 2010. Consumer acceptance of sms advertising: a study of american and turkish consumers. *International Journal of Mobile Communications* 8(4): 392–410.
- Oliva, Jesus; Serrano, J. Ignacio; Del Castillo, M. Dolores; Iglesias, Ángel (2013). A SMS normalization system integrating multiple grammatical resources. *Natural Language Engineering* 19: 121–141.
- Panckhurst, Rachel. 1997. « La communication médiatisée par ordinateur ou la communication médiée par ordinateur? », *Terminologies nouvelles* 17 : 56–58.
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In Arnavielle, T., (ed.) : *Polyphonies*, pour Michelle Lanvin: 33–52. Université Paul-Valéry Montpellier 3, Montpellier.
- Schneider-Hufschmidt, Matthias. 2005. Usability Issues of Sending Text Messages. In Harper, R., Palen, L. A.; Taylor, A.S. (eds). *The Inside Text*: 223–236. Springer, Dordrecht.
- Stark, Elisabeth. 2011. La morphosyntaxe dans les SMS suisses francophones : Le marquage de l'accord sujet — verbe conjugué. *Linguistik online* 48(4): 35–47.

- Tagg, Caroline. 2009. *A corpus analysis of SMS text messaging*. PhD thesis, University of Birmingham.
- Thurlow, Crispin. 2003. Generation txt ? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1.
- Turoff, Murray; Starr Roxanne Hiltz. 1977. *Meeting Through Your Computer*. IEEE Spectrum, May.
- Vold Lexander, Kristin. 2007. Langues et SMS au Sénégal. Le cas des étudiants de Dakar. In Gerbault, J. (ed.). *La langue du cyberspace : de la diversité aux normes*: 59–67. L'Harmattan, Paris.
- Yvon, François. 2008. *Reorthography of SMS messages*. Research Report LIMSI/CNRS, 2008.
- Walkowska, Justyna. 2009. Gathering and Analysis of a Corpus of Polish SMS Dialogues. In Kłopotek, M.A.; Przepiórkowski, A.; Wierchoń, S.T.; Trojanowski, K. (eds). *Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems*: 145–157, Academic Publishing House EXIT, Warsaw.
- Wei, Ran; Hao Xiaoming; Ji Pan. 2010. Examining user behavioral response to sms ads: Implications for the evolution of the mobile phone as a bona-fide medium. *Telematics and Informatics* 27(1): 32–41.

Seek&Hide

Anonymising a French SMS corpus using natural language processing techniques

Pierre Accorsi*, Namrata Patel*, Cédric Lopez**,
Rachel Panckhurst***, and Mathieu Roche****

*Université Montpellier 2, LIRMM, CNRS / **Objet Direct — VISEO /

***Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3 /

****TETIS, Cirad, Irstea, AgroParisTech and LIRMM, CNRS, Université
Montpellier 2

Introduction

In a short span of years, SMS (Short Message Service) communication has become a veritable social phenomenon (over 60 billion text messages are sent every year!¹). Although numerous scientific studies (namely in the fields of linguistics, sociology, psychology, mass communication, etc.) have been conducted on this recent form of communication, there remains a general gap in our accumulated knowledge of the subject. This is mainly due to the fact that researchers have limited access to suitable data for their studies. Typically, they require large volumes of authentic data for their work to be significant.

The international project *sms4science*² (coordinated by CENTAL³) aims at building and studying precisely such a body of data by collecting authentic text messages from different parts of the world. In this context, over 90,000 authentic text messages in French were collected across a three-month period (spanning 15 September to 15 December 2011), by a group of academics in the Languedoc-Roussillon region of France, under the following programme: *sud4science LR*.

However, in order to publish or distribute these resources, researchers are legally required to meticulously remove all traces of identification from each SMS.

1. According to the ARCEP (Autorité de régulation des communications électroniques et des postes).

2. <http://www.sms4science.org/>

3. <http://www.uclouvain.be/cental.html>

Furthermore, there are cases in which people are explicitly mentioned in a text message: donors (people who have agreed to donate their SMSes for the study), recipients (receiving the text message), third parties (including researchers conducting the study). The process of SMS anonymisation is therefore deemed essential on ethical grounds and is a crucial step before any form of public distribution of the corpus.

Rendering a document anonymous does not merely imply the removal of all proper names that it contains. As L. Sweeney *et al.* (2002) point out, 87% of the U.S. population can be uniquely identified on the basis of their postcode, date of birth and sex, while 97% of the population of Cambridge, Massachusetts, can be identified using only their postcode and date of birth. It thus seems fitting that anonymisation be accomplished using natural language processing techniques.

To this end, a piece of software using human-aided automatic text processing, *Seek&Hide*,⁴ was designed. This article presents the system of anonymisation devised for the software, consisting of an automated process tailored to the requirements of our *sud4science LR* corpus, assisted by a semi-automatic process. The current version of *Seek&Hide* focusses on the anonymisation of first names, as this constitutes the major part of the most difficult words processed. For instance, first name spellings can vary even when they refer to the same individual: *Nicolas, Nico, Nicooo, Niko, Nicoco, Nyko, etc.* Other elements that require anonymisation are processed in a semi-automatic phase, aided by an expert annotator (§2).

We begin by introducing the distinctive aspects of our work by looking at pre-existing anonymisation techniques (§1). After outlining the way in which *Seek&Hide* functions, we examine the problems (specific to our corpus) along with the solutions that *Seek&Hide* proposes to remedy them (§2). To conclude our study, we present and discuss the results of tests that were conducted on a sample in order to evaluate *Seek&Hide*'s performance (§3).

1. Literature review

Anonymisation is vital when one seeks to mask an individual's identity. For example, this is essential before the distribution of court orders pertaining to children, juvenile delinquents, victims of sexual harassment, etc. (L. Plamondon *et al.* 2004), or when one needs to put together a medical corpus (C. Grouin *et al.* 2009).

In the medical field, it is customary to resort to automatic anonymisation techniques using rules and medical dictionaries in order to process the most common cases (L. Plamondon *et al.* 2004; L. Sweeney 1996; E. Aramaki *et al.* 2006; C.

4. Not "Hide and Seek", but "Seek and Hide": with this tool, we seek to hide words that are to be anonymised.

Grouin *et al.* 2009; J. Gardner *et al.* 2010). These systems primarily aim at the automatic recognition of names, dates, places and other elements which could lead to the identification of people covered by publication restrictions.

We agree with C. Reffay *et al.* (2012) that the process of anonymisation cannot be entirely automated. Their work focusses on the creation of an interface by which the researcher can identify personal data and decide whether or not to render it anonymous. Given the size of our *sud4science LR* corpus (over 90,000 SMSes), an automated procedure considerably benefits the annotator, as we shall show during the course of this article. The distinguishing feature in our approach, as put forward by C. Fairon and J.R. Klein (2010), is that we take into consideration the numerous linguistic particularities of the forms used in SMS writing.

In such situations, the identity markers that need to be anonymised are names made up by the senders themselves, which often become cultural footprints (nicknames, diminutives) (Chr. Reffay and Ph. Teutsch 2007), *i.e.* in the following text messages, “Coco” and “cece” require anonymisation.

Coco est pas la ! Éva non plus ! [...]
Coucou mon cece ! J'espere [...]

In the following section we propose an anonymisation technique which is adapted to the demands of text messages.

2. Seek&Hide: A system for anonymising a French corpus

As stated in the previous section, our approach to anonymise/de-identify the *sud4science LR* corpus of French text messages is to adopt a two-phase procedure.

To this effect, *Seek&Hide*, the SMS processing tool developed for this task, functions in the following two phases:

Automatic Phase: automatic processing of the corpus and labelling of the information which needs to be rendered anonymous.

Semi-automatic Phase: manual processing of the information that could not be treated during the automatic phase followed by overall validation and error correction via a web interface.

This two-phase procedure ensures the dependability of the system: the combined use of Natural Language Processing (NLP) techniques and human evaluation helps minimise computer as well as human errors, greatly improving the overall result.

In terms of software development, *Seek&Hide* has been designed as a dynamic web page, in keeping with the idea of simplicity and ease of use. All its processes

have thus been written in PHP (a general server-side scripting language⁵), in order to maintain its integrity.

Let us now take a deeper look into the workings of this system by considering each of its processes individually.

2.1 The automatic phase

We begin with the first phase of the system: the automatic phase.

2.1.1 Purpose

The main purpose of the automatic phase is to process the *sud4science LR* corpus using (a) dictionaries as reference material and (b) word-processing techniques so as to identify and, if required, hide words that have to be anonymised. This constitutes the preliminary treatment of the *sud4science LR* corpus. Text messages that are processed by this phase undergo the following transformation, the details of which are given in the next paragraph:

“Coucou Patrice, ça va?” => “Coucou <PRE_7_17316>, ça va?”

Anonymisation codes When a word is anonymised, it is replaced by a code which follows the following format: <[Tag]_[#characters]_[cross-reference]> where “Tag” indicates the type of the word (e.g. First name, Last name). Thus, “Patrice” is replaced by < PRE_7_17316 > where:

- PRE => First name (prénom)
- 7 => number of characters in “Patrice”
- 17316 => “Patrice”’s ID in the dictionary of first names

On the basis of this format, (a) first names can be distinguished from last names, (b) the number of characters in the SMS is preserved (for statistical purposes) and (c) each word that is anonymised can be retraced back to its origin in the dictionary, thanks to its ID.

2.1.2 Process: The Seek&Hide Algorithm

Having seen what *Seek&Hide* does in its automatic phase, let us now find out how this is done. *Seek&Hide* operates in a three-step procedure:

- Pre-processing
- Identification of the words to be anonymised
- Processing of the words to be anonymised

5. <http://www.php.net>

2.1.2.1 Pre-processing. Each SMS in the corpus, in its raw state, is basically just a string of characters. In order for *Seek&Hide* to make any sense out of this data, it needs to break the string of characters into words. This is the pre-processing phase, called “Tokenisation”.

Once tokenised, the SMS becomes a coherent sentence: a series of identifiable words. SMS tokenisation is a complex process (R. Beaufort *et al.*, 2010). For an anonymisation task that does not need a precise analysis of message content, such as ours, we consider that the simple use of a “space” as separator for tokenisation is satisfactory.

2.1.2.2 Identification. In this phase of the automatic process, *Seek&Hide* uses a technique of identification which uses specific kinds of dictionaries to analyse each word of an SMS. The idea behind this is simple: each word of an SMS can either be classified as “To anonymise” or as “Nothing to anonymise”. We thus use two kinds of dictionaries corresponding to this classification, distinguishing them by “Dictionary” and “Anti-dictionary” on the basis of their content (pertaining to the task of anonymisation):

- The “Dictionary” contains words that need to be anonymised.
- The “Anti-dictionary” contains words that do not require anonymisation.

The following list shows the different resources used as reference material to identify the words in the text messages:

Dictionary: Dictionary of first names

Anti-dictionaries:

- Dictionary of inflected forms of the French language (Lexique des Formes Fléchies du Français, LEFFF)⁶
- Dictionary of some forms used in SMS writing
- Dictionary of places

Each word is then labelled according to its presence or absence, in the dictionaries used by *Seek&Hide* (cf. Table 1). As mentioned in §1, here we take into account the numerous linguistic particularities of the forms used in SMS writing by using heuristic methods. In the next section (§2.1.3), we look more closely into this aspect providing details on each of the solutions that *Seek&Hide* proposes.

2.1.2.3 Processing. By now, the corpus is ready for a first round of anonymisation: the automatic anonymisation. Words are processed according to their labels and are thus (a) anonymised, (b) ignored, (c) highlighted. Words that could not be

6. <http://www.labri.fr/perso/clement/lefff/>

identified in a dictionary, and words that were identified in both types of dictionaries, are highlighted.

These will be processed via a web-interface during the semi-automatic phase of the system (cf. §2.2).

Table 1 summarises this treatment by giving the range of possible cases encountered.

As can be seen, “Cédric” is anonymised because it is identified only in the dictionary (of first and last names). Similarly, “crayon” is ignored as it is identified only in the anti-dictionary (LEFFF). “Pierre” and “Namrata” are problematic: “Pierre” is ambiguous as it belongs to both, the dictionary and the anti-dictionary.

Table 1. The automatic phase

Word processed	In dictionary?	In anti-dictionary?	Label	Treatment
Cédric	yes	no	Dictionary	Automatically Anonymised
Crayon	no	yes	Anti-dictionary	Ignored (not to be anonymised)
Pierre	yes	yes	Ambiguous	Highlighted (candidate for the semi-automatic phase)
Namrata	no	no	Unknown	Highlighted (candidate for the semi-automatic phase)

“Namrata” is unknown as it belongs to neither of the dictionaries. These two words are consequently highlighted for further processing.

This summarises the basic function of the automatic phase of the *Seek&Hide* system. However, before moving on to the next phase, we would like to dedicate the following section to examining the problems encountered during this phase, and the methods we used to solve them.

2.1.3 *Heuristics for SMS processing*

The tricky part of the automatic processing of the text messages lies in the dictionary word search procedure. It proves to be inefficient in certain cases, namely the following:

- Misspelled words: ex.: *surment* (instead of *sûrement*)
- Words written without their accents: ex.: *desole* (instead of *désolé*)
- Words with incorrect accents: ex.: *dèsolè* (instead of *désolé*)
- Letter repetitions: ex.: *nicoooooIIlaassss* (instead of *nicolas*)
- Onomatopoeias: ex.: *mouhahaha*
- Omission of the apostrophe: ex.: *jexplique* (instead of *j’explique*)
- Concatenation: ex.: *jtaime* (instead of *je t’aime*)

Table 2 describes heuristic solutions devised to improve the accuracy of word search in the dictionaries:

Of the algorithms listed, LRepet is the most complex one. It also proves to be the most interesting in terms of NLP. Let us therefore take a closer look at it.

The LRepet Algorithm:

Step 1: Reduce every word so that it contains, at most, double letters. (Ex. *Nicoooooolllaassss => Nicoollaass*). This is justified since no word contains more than two consecutive identical letters in French. Also, this considerably reduces the search span in Step 2.

Table 2. Heuristic solutions

Algorithm	Problem	Solution
WWoutA	Words written without their accents (<i>desole</i>)	Perform an “accent-insensitive” word-search
WWithA	Words with incorrect accents (<i>dèsolè</i>)	Remove the accents from the word and then perform the search
OmiA	Omission of the apostrophe (<i>jexpli-que</i>) and Concatenation (<i>jtaime</i>)	Identify and remove prefixes such as <i>jt</i> , <i>jl</i> , <i>j</i> , <i>etc.</i> , then perform the word-search
SRepet	Onomatopoeias (<i>mouhahaha</i>)	Detect repetitions of sub-strings such as <i>ha</i> , <i>hé</i>
LRepet	Letter repetitions (<i>Nicoooooolllaassss</i>)	Delete consecutive identical letters and search

Step 2: Perform a dictionary search for each possible combination of double letters. Examining Table 3, we can see that in our example (Ex. *Nicoooooolllaassss => Nicoollaass*), 16 cases are to be processed. Here we must note that the occurrence of four double letters in a single word is extremely rare for the majority of cases in our actual data.

Table 3. LRepet — Step 2: Search cases for Nicoooooolllaassss

# Double-Letters	# Possible Cases	Cases
4	1	<i>Nicoollaass</i>
3	4	<i>Nicollaass</i> , <i>Nicoolaass</i> , <i>Nicoollass</i> , <i>Nicoollaas</i>
2	6	<i>Nicoollas</i> , <i>Nicoolaas</i> , <i>Nicoolass</i> , <i>Nicollaas</i> , <i>Nicollass</i> , <i>Nicolaass</i>
1	4	<i>Nicoolas</i> , <i>Nicollas</i> , <i>Nicolaas</i> , <i>Nicolass</i>
0	1	<i>Nicolas</i>

In order to keep the search space as small as possible, this algorithm is used sparingly: it is performed only on the words that are unknown (that exist in none of the dictionaries). In future work, we would like to further reduce the complexity of this algorithm by using linguistic rules that are specific to languages (*i.e.*, in cases where consecutive double letters might be impossible).

This algorithm also helps us to identify incorrectly spelled words, as shown in the following example:

Ex. “*Philippe*” (*correct*) identified from “*Phillippe*” (*incorrect*).

We are now equipped to fully appreciate the importance of the next phase of the system: the semi-automatic phase.

2.2. The semi-automatic phase

The semi-automatic phase in our system was designed to ensure that every SMS in the corpus is correctly anonymised. Specifically, this means carrying the task out in a two-fold process: (a) Evaluating the results of the automatic phase, the details of which are presented in the following section (§3) and (b) Manually processing the text messages termed ambiguous (*i.e.* texts containing at least one word which is either unknown or found in both the dictionary of first names as well as the ‘anti-dictionaries’).

The *sud4science* LR corpus of about 90,000 text messages contains $\approx 27\%$ which need to be processed manually.

The term ‘semi-automatic’ means that *Seek&Hide* assists the human annotator by highlighting (bold and italic) words that require intervention (*cf.* Figure 1), namely those that belong to the category described as “highlighted” (*cf.* Table 1) in the previous section (*e.g.* “cece”, “bebou”, “surment”). In such cases, the annotator can disambiguate the words only by reviewing them in their context, as is reflected in the word “cece” in the following SMS gleaned from our corpus (also seen in Figure 1 — SMS #8):

Hey hey ! J'te souhaite un tres joyeux n'anniversaire! Profite bien de ta journee ! Le cece te souhaite aussi un tres joyeux anniv' ! Bisoux

In the text messages that are processed during this phase, the words that were present only in the dictionary of first names bear the tag < PRE >. If this word needs to be stripped of its anonymisation tag, the annotator has the possibility to do so with a single click. Note that this tag does not fully respect the SMS anonymisation code specifications (*cf.* §2.1.2.3). This is merely to avoid cluttering the screen with incomprehensible codes; the final corpus that is generated at the end of this phase contains correctly formatted tags.



Figure 1. Screenshot of the web interface



Figure 2. Popup Window containing options for anonymisation tags

Further, when the annotator clicks on a word that requires anonymisation (whether highlighted on the screen or not), a window pops up. This window, (cf. Figure 2) lists options for the anonymisation tag:

- First Name (PRE)
- Last Name (NOM)
- Nickname (SUR)
- Address (ADR)
- Place (LIE)
- Tel. Num. (TEL)
- Code (COD)
- URL (URL)

- Brand Name (MAR)
- E-mail (MEL)
- Other

Once chosen, the tag is added to the appropriate word, containing within it an explicit mention of “USER” (e.g. < PRE_USER >). Moreover, this additional information is stored in specific anti-dictionaries pertaining solely to the users of *Seek&Hide*.

Having completed modifications/verifications (additions/deletions of anonymisation tags) for each SMS, the annotator can finalise them by clicking on the button labelled “Je valide” (in English, “I confirm”) (cf. Figure 3). (S)He can also choose to ignore by clicking on the button labelled “Je ne sais pas” (in English, “I don’t know”), and/or add a comment by clicking on the button labelled “Commenter” (in English, “Comment”).

Again, we call attention to the fact that this article focusses on the most complex part of the anonymisation process: that of the processing of first names. Other anonymisation tasks are accomplished using regular expressions to identify the appropriate words; for example, e-mail addresses, telephone numbers, and URLs.

The welcome page for our piece of software, shown in Figure 4, displays all its features. In the first box (Anonymisation), we distinguish between “semi-automatic anonymisation” and “automatic anonymisation”. Note that the latter is perceived

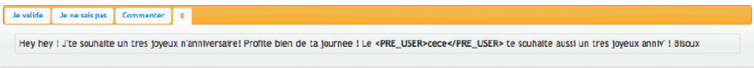


Figure 3. Zoom on a manually annotated SMS

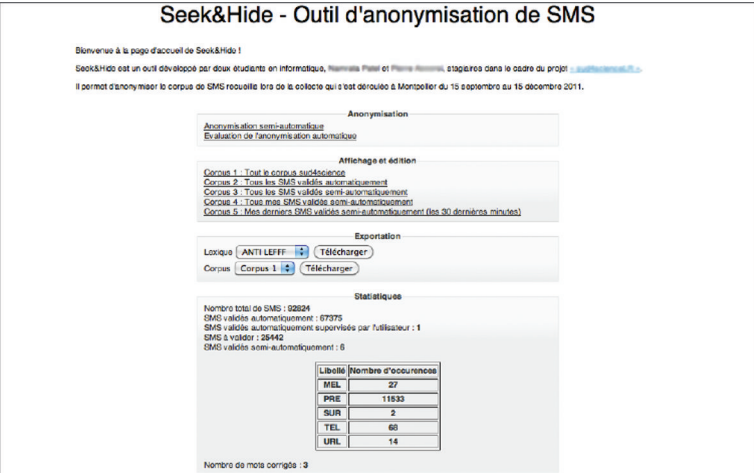


Figure 4. Welcome page for *Seek&Hide*

as an evaluation of the automatic phase of our system, based on the results of the validation performed by an expert. Additionally, using the options in the second box (Affichage et édition: in English, ‘view and edit’), the expert can also display/edit (and therefore correct) the text messages validated by other experts. Using the drop-down lists in the third box (Exportation: in English, ‘export’), the expert can download this processed/validated data along with the various dictionaries used by *Seek&Hide* during its automatic phase. The last box (Statistiques: In English, ‘statistics’) displays numerous statistics on the words and the text messages that are processed.

Thus, even before the semi-automatic processing phase to be conducted during subsequent months, we have the following information at our disposal:

- Total number of text messages: **92,824⁷**
- Text messages automatically processed: **67,376**
- Text messages automatically processed, validated by an expert: **0**
- Text messages left to validate: **25,448**
- Text messages semi-automatically validated: **0**

Furthermore, we have the following information concerning the various kinds of words anonymised, using dictionaries and regular expressions:

- **11,532** first names
- **27** e-mail addresses
- **68** telephone numbers
- **14** URLs

Lastly, the number of modifications carried out (number of words corrected) are provided as indicative values.

With that, we come to the end of our presentation of the workings of *Seek&Hide*. In the next part of this article, we proceed to an assessment of this tool.

3. Experiments

3.1 Experimental protocol

Seek&Hide was tested on a sample of the *sud4science LR* corpus containing 23,055 SMSes that were manually tagged as “To anonymise (TA)” or “Nothing to anonymise

7. This includes duplicate text messages. When eliminated from the corpus, the total number of text messages is 88,683. The number of automatically processed text messages is 67,376. The number of text messages left to validate is 24,955.

(NTA)” by a student annotator. During the acquisition of the *sud4science LR* corpus, a fourth year student was employed for a three-month internship, in order to read and label the incoming messages to make sure they respected certain rules and regulations. Out of the 23,055 SMSes in our sample, 90.7% (*i.e.* 20,913 SMSes) were noted by him as NTA and 9.3% (*i.e.* 2,142 SMSes) as TA.

In the following section, we propose a method of evaluation whereby *Seek&Hide*’s results on the sample are compared with those of the student annotator.

The evaluation consists of checking whether an SMS (as a whole) has to be anonymised or not. Note that our tool further highlights those individual parts of the SMS which it deems require anonymisation, for the benefit of the annotator, in order to simplify his/her task.

3.2 Global analysis of the *Seek&Hide* results

Table 4 presents the SMS distribution in the sample according to three categories: Those tagged TA, those tagged NTA, and those left untagged. The untagged label corresponds to the text messages that our automatic system cannot tag (TA or NTA) because they contain ambiguous and/or unknown words. These will be processed via the web-interface for the semi-automatic phase of our tool.

Table 4. Results of the *Seek&Hide* system on the sample analysed by the student annotator.

Sample Processed	Tag: TA	Tag: NTA	Untagged	Total
By Seek&Hide	1,089	13,963	8,003	23,055
	4.72%	60.57%	34.71%	

We note that our system returns results for 65.3% of the sample corpus (*i.e.* 15,052 SMSes with TA and NTA tags). The other part of the corpus (34.7 % of the corpus left untagged) has to be processed by the semi-automatic system. This point will be further discussed in §4.

3.2.1 *Method of evaluation*

In this section we focus our analysis on the evaluation of the 15,052 SMSes automatically processed by *Seek&Hide* as NTA and TA. In this context, the following confusion matrices (see Tables 5 and 6) show a more detailed analysis of the results obtained.

Definitions of TP (true positive), FP (false positive), TN (true negative) and FN (false negative) with respect to the tag ‘NTA’ (Table 5) are provided in the following paragraph. Those with respect to the tag ‘TA’ (Table 6) can be inferred accordingly.

Table 5. Confusion Matrix with respect to the tag ‘NTA’

NTA = “correct”	Student-Annotator: NTA	Student-Annotator: TA
Seek&Hide: NTA	13,904	59
Seek&Hide: TA	413	676

Table 6. Confusion Matrix with respect to the tag ‘TA’

TA = “correct”	Student-Annotator: TA	Student-Annotator: NTA
Seek&Hide: TA	676	413
Seek&Hide: NTA	59	13,904

The top left box indicates the number of TP (*i.e.* the 13,904 SMSes correctly classified by the application as ‘NTA’), the top right, FN (*i.e.* the 59 SMSes classified as ‘NTA’ by *Seek&Hide* and ‘TA’ by the student-annotator), the bottom left, FP (*i.e.* the 413 SMSes classified as ‘TA’ by *Seek&Hide* and ‘NTA’ by the student-annotator), and the bottom right, TN (*i.e.* the 676 SMSes correctly classified as ‘TA’). Note that correct and incorrect terms are based on the tags specified by the student-annotator.

A deeper analysis of Table 5 shows that *Seek&Hide* predicts 13,963 SMSes (*i.e.* first line of Table 5: 13,904+59) that do not need anonymisation (tagged NTA). Of these, only 59 SMSes do not correspond to the manual evaluation performed by the student annotator. This shows that the NTA-tagging performed by *Seek&Hide* is very efficient. However, the *Seek&Hide* prediction of text messages that require anonymisation (*i.e.* second line of Table 5) is not as good, as only 676 of 1,089 SMSes are relevant.

The global result of our automatic system can be given by computing the *Accuracy*. The accuracy, based on the calculus of the diagonal of the confusion matrix, is given by the following formula [1]:

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad [1]$$

In our case we obtain a value of 0.96⁸ for the accuracy. This score validates the relevance of our methods in predicting text messages that may or may not require anonymisation.

We can explain the 4% error by looking at the following situations:

8. $\frac{13,904 + 676}{13,904 + 676 + 413 + 59}$

1. Forms of SMS writing which are not listed in the anti-dictionaries (ex. “sora” is (a) a female name and (b) a future tense form of the French verb savoir “saura”) (**false positive**)
2. Short forms of ambiguous words (ex. “just” is short for (a) “Juste” — a male name and (b) “juste” — the French adjective/noun/adverb) (**false negative**)
3. Words from foreign languages which are first names (ex. “will” is (a) a male name and (b) a future tense form in English) (**false positive**)
4. First names that are also brand names (ex. “mercédès” is (a) a female name and (b) the German car brand) (**false positive**)

We also evaluated these results using measures from the text-mining domain (cf. Table 7). This greatly improves our analysis as it provides further observations on Seek&Hide’s performance.

Table 7. Performance measures

	Precision	Recall	F-Measure
NTA	0.996	0.971	0.983
TA	0.621	0.920	0.741

These measures correspond to:

Precision (Formula [2])

$$Precision = \frac{TP}{TP + FP} \quad [2]$$

A measure which evaluates the relevance of the results obtained. A high precision thus implies that the system returned more relevant results than irrelevant ones.

Recall (Formula [3])

$$Recall = \frac{TP}{TP + FN} \quad [3]$$

A measure which evaluates the number of relevant results obtained. A high recall thus implies that a large part of relevant elements are returned by the system.

F-measure (Formula [4])

$$F\text{-measure} = 2 \cdot \frac{P \cdot R}{P + R} \quad [4]$$

A measure that combines the values of “Precision” (P) and “Recall” (R) to provide a balanced evaluation.

Table 7 shows the precision, recall and F-measure calculated for each of the confusion matrices (Tables 5 and 6). The results confirm the good performance of our system from an overall point of view (F-measure for both ‘NTA’ and ‘TA’). The precision of the ‘TA’ tag still needs to be improved in our future work.

Having obtained such positive evaluations for *Seek&Hide*, we decided to further examine the contributing factors in the *Seek&Hide* algorithm. In the following section, we provide a detailed analysis of the heuristics (§2.1.3) used during the automatic processing of the *sud4science LR* corpus.

3.3 Evaluation of the *Seek&Hide* heuristics

Table 8 shows quantitative results on the contributions of the different heuristic algorithms (*i.e.* number of SMSes detected by running of each of these algorithms) on the entire corpus as well as the test sample. These results reveal how crucial it is to take into account the various aspects (e.g. accents, apostrophes and repeated characters) that make the words contained in these SMSes particularly problematic. Indeed, they are found in nearly 40,000 SMSes.

A qualitative evaluation of these heuristics was also conducted on a portion (*i.e.* 1,800 SMSes) of the test sample. This particular kind of evaluation reveals the

Table 8. Quantitative evaluation of the heuristic algorithms

Algorithm	# SMSes detected in corpus	# SMSes detected in test sample
WWoutA	3,970	820
WWWithA	19,791	5,198
OmiA	9,261	2,395
SRepet	940	205
LRepet	10,428	2,399

Table 9. Qualitative evaluation of the heuristic algorithms

Algorithm	# SMSes detected in sample	# SMSes correctly detected in sample	% of correct results
WWoutA	113	110	97.3
WWWithA	996	989	99.3
OmiA	481	466	96.9
SRepet	44	44	100.0
LRepet	2,230	2,163	97.0

rate of correct results obtained by our heuristics. This is shown in Table 9. The high percentage of correct results obtained (between 96.9% and 100% according to the applied algorithms) shows that these algorithms are indeed an effective aid to our system.

Conclusion and future work

The system proposed in this article performs the anonymisation/de-identification of a corpus. To this end, it uses (a) a dictionary of first names and (b) anti-dictionaries (of ordinary language and of some forms of SMS writing) to identify the words that require anonymisation. Note that the adopted principle is sufficiently generic for it to be adapted to various types of corpus, irrespective of their language.

In its automatic phase, the system processes over 70% of the corpus. This corresponds to the unambiguous text messages present in it: those that contain words that are neither unknown, nor ambiguous (found in both the dictionary as well as the anti-dictionaries). A comparative analysis of its performance, based on the manual evaluation of a significant albeit small portion of the corpus (*i.e.* 23,055 SMSes), yielded positive results on 96% of the text messages processed (whether considered to be anonymised or not to be anonymised).

The basic principle behind the processing tool *Seek&Hide* is to highlight words that require manual verification. These words, displayed within their context, provide valuable assistance to the annotator. As a result, his/her task is greatly simplified: on average, an expert takes 180 seconds to process 20 text messages using our tool.⁹

As future work, two students will perform the task of anonymisation. A thorough analysis on their part will allow us to improve our techniques and enrich our dictionaries. We will then be able to analyse the contributions of the different dictionaries used and examine words that emerge from both the automatic and the semi-automatic process. This analysis will enable us to further our analysis of the different phases of the *Seek&Hide* processing tool. Moreover, we would like to apply this tool and its associated algorithms to other types of data (e.g. medical data) that require anonymisation/de-identification. Finally, the fully anonymised corpus (containing the interns' annotations) could be used as a training set to help us explore the possibilities of supervised learning, thus allowing our system to learn the rules of anonymisation automatically.

9. Taking this into account ($\approx 9\text{s/SMS}$), the task of anonymisation should amount to around two months' full-time work. As future interns are intended to take this up for a two month period, it seems plausible that they should accomplish it in the given time with the help of *Seek&Hide*.

Acknowledgements

The *sud4science LR* project is part of a vast international SMS data collection project, entitled *sms4science* (<http://www.sms4science.org>), and started at the CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium) in 2004. Our work is supported by the MSH-M (Maison des Sciences de l'Homme de Montpellier) and the DGLFLF (Délégation générale à la langue française et aux langues de France).

References

- Aramaki, Eiji; Takeshi Imai; Kengo Miyo; Kazuhiko Ohe. 2006. Automatic deidentification by using sentence features and label consistency. In *I2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 10–11.
- Beaufort, Richard, Sophie Roekhaut; Louise-Amélie Cougnon; Cédric Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 770–779, Uppsala, Sweden: Association for Computational Linguistics.
- Fairon, Cédric; Jean René Klein. 2010. Les écritures et graphies inventives des SMS face aux graphies normées. *Le Français aujourd'hui* 3: 113–122.
- Fairon, Cédric; Jean René Klein; Paumier Sébastien. 2006. *SMS pour la science: Corpus de 30.000 SMS et logiciel de consultation*. Louvain-la-Neuve: Presses universitaires de Louvain, Manuel+CD-Rom.
- Gardner, James; Li Xiong; Fusheng Wang; Andrew Post; Joel Saltz; Tyrone Grandison. 2010. An evaluation of feature sets and sampling techniques for de-identification of medical records. In *Proceedings of the 1st ACM International Health Informatics Symposium*, 183–190, ACM.
- Grouin, Cyril; Arnaud Rosier; Olivier Dameron; Pierre Zweigenbaum. 2009. Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. In *Risques, technologies de l'information pour les pratiques médicales*, M. Fieschi, P. Staccini, O. Bouhaddou, C. Lovis, 23–34, Paris: Springer.
- Plamondon, Luc; Guy Lapalme; Frédéric Pelletier. 2004. Anonymisation de décisions de justice. *TALN'04*.
- Reffay, Christophe; François-Marie Blondel; Emmanuel Giguët. 2012. Stratégies pour l'anonymisation systématique d'un corpus d'interactions plurilingues. In *Actes du colloque IC2012. Intercompréhension: compétences plurielles, corpus, intégration*, 1–21.
- Reffay, Christophe; Philippe Teutsch. 2007. Anonymisation de corpus réutilisables Masquer l'identité sans altérer l'analyse des interactions. *EIAH'2007 Conference: Environnements Informatiques pour l'Apprentissage Humain*, accepted as a poster.
- Sweeney, Latanya. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA Annual Fall Symposium*, 333, American Medical Informatics Association.
- Sweeney, Latanya. 2002. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems* 10(5): 557–570.

Summary

This article presents the system *Seek&Hide*, a text message processing tool developed for the *sud4science LR* (<http://www.sud4science.org/>) project. It performs the anonymisation/de-identification of a corpus. At present, it has been used to anonymise the *sud4science LR* corpus of French text messages collected during the project. This is done in two phases. In the first phase, it automatically processes over 70% of the corpus. The rest of the corpus is processed in the second phase, aided by an expert annotator via a web interface specifically designed to simplify the task.

SMS experience and textisms in young adolescents

Presentation of a longitudinally collected corpus

Josie Bernicot*, Olga Volckaert-Legrier**, Antonine Goumi***,
and Alain Bert-Erboul*

*Université de Poitiers — CNRS / **Université Toulouse II — Le Mirail /

***Université Paris Ouest — Nanterre La Défense

Introduction

Text messages (SMS) have ‘taken over’ daily life, bringing to the fore written forms which would have been unimaginable 20 years ago (for example, “*the gr8 db8*” or, in French, “*1 pw1 sr la kestion*”). Studies carried out in the United States (Pew Internet and American Life Project, A. Lenhart, R. Ling, S. Campbell and K. Purcell 2010) and in France (CREDOC, L. Bigot and P. Crouette 2011) confirm this feeling, emphasizing the young age of the texters who use SMSes as a core part of their social relations. In France in 2011, 82% of 12 to 17 year-olds had a mobile phone, and 99% used their mobile phone to send SMSes. Over the past 10 years, the number of studies focusing on this new mode of communication has continued to grow. Consistent information in several languages is now becoming available concerning message length, speaking turns, openings, closings, spelling changes with regard to the traditional written code (textisms), gender differences (cf. the summary by C. Thurlow and M. Poff 2013) as well as smileys (C. Tosell, P. Kortum, C. Shepard, L. Barg-Walkow, A. Rahmati and M. Zhong 2011). The scientific questioning with regard to SMSes has led to the construction of databases of messages produced in natural interactions. C. Fairon, J.R. Klein and S. Paumier’s (2006) “sms4science” project resulted in a database containing 30,000 SMSes written by 2,436 French-speaking Belgian ‘informants’ between the ages of 12 and 73.

The phenomenon of SMS use has developed so swiftly that it has not allowed scientific studies to easily grasp how texters learn this new written code. The study presented in this paper focuses on that point, by analysing the SMSes written by young adolescents between the ages of 11 and 12 during their first year of SMS use. New SMS users are confronted with a completely new interaction situation

in which they are not immediately able to master all the rules. From a pragmatic point of view, a fundamental rule is the linking of the linguistic forms of utterances with the characteristics of the interaction situation (J. Austin 1962; J. Bernicot 1994; J. Bernicot and A. Mahrokhian 1989; J. Bernicot, J. Comeau and H. Feider 1994; J. Bernicot, V. Laval and S. Chaminaud 2007; P. Grice 1975; I. Noveck and D. Sperber 2004; J. Searle 1969; J. Verschueren 1999). The concepts of register (D. Ravid and L. Tolchinsky 2002) or “language variety” (D. Crystal 2001) enable the linguistic specificities of an interaction situation to be determined. It was from this angle that the current study examined a corpus of 4,524 SMSes sent by young adolescents in daily life situations and collected in a longitudinal manner over a period of 12 months. The objective was to show how beginning texters became able to master the structural characteristics of SMSes and, in particular, their orthographic forms (textisms). The summary of research already carried out in this area is presented below.

1. Orthographic forms of SMSes (textisms) in young adolescents

Early research took into account the existence of abbreviations in SMSes and calculated the percentage of SMSes with abbreviations or the percentage of participants using abbreviations. In R. Ling’s (2005) corpus, 6% of the messages were written with abbreviations: girls between the ages of 13 and 15 were the greatest users of abbreviations (20% of their messages contained them).

More recent research has taken into account the density of textisms as a reference index. A textism is defined as a change in a word’s orthographic form as compared to traditional writing. For each message, the density of textisms is equal to the number of changes divided by the total number of words in the message. Studies of children or young adolescents (9–12 years old) are rare and do not point out any specific forms of textisms for this age group. The classifications used are derived from those of C. Thurlow and A. Brown (2003) which include the following 10 categories for the English language.

- Shortenings (bro for brother)
- Contractions (gd for good)
- G-clippings (goin for going)
- Other clippings (hav for have)
- Acronyms (BFPO for British Forces Posted Overseas)
- Initialisms (V for very)
- Letter/number homophones (2moro for tomorrow)
- Misspellings (cuming for coming)

- Non-conventional spellings (fone for phone)
- Accent stylizations (afta for after)

B. Plester, C. Wood and P. Joshi (2009) asked 88 British children (mean age: 10 years and 7 months) to write out the messages that they had written on their mobile phones in 10 daily life situations (e.g.: telling a friend in class about being late due to a bus not stopping, or letting a mother know that she forgot to buy dog food). The average age at which the first mobile phone was acquired was 9 and the participants therefore had approximately one and a half years of experience. The proportion of textisms was .34, with a difference between girls and boys (respectively .38 for girls and .28 for boys). The authors also studied the kinds of textisms and showed that two major categories stood out: simplifications (accent stylizations) which conserve the pronunciation of the spoken language (ex.: afta for after) and replacements of letters with a number homophone (e.g.: 2moro for tomorrow).

P. Plester, M. K. Lerkkanen, L. Linjama, H. Rasku-Puttonen and K. Littleton (2011) analysed the SMSes of 65 young Finnish adolescents between the ages of 9 and 11 (mean age: 10 years and 7 months). The children reported having received their first mobile phone at the age of 7 years and 3 months, and therefore they are considered as having more than 3 years of texting experience. The participants were asked to copy down the SMSes they had sent over the weekend and give them to the researcher when they arrived back at school. The average proportion of textisms was .48 and simplifications (accent stylizations) conserving the pronunciation of spoken Finnish dominated all the other categories. The replacements of letters with a number homophone were never used. The most-used types of textisms therefore vary from one language to another.

N. Kemp and C. Bushnell (2011) asked 86 children between the ages of 10 and 12 (mean age: 11 years and 6 months) to write a message in SMS language on a mobile phone. The message was dictated by a researcher and the participants were to imagine that they were sending it to a friend (ex: When will we see you tonight? Because someone left a message about your friend being sick. Are you sick too?). When the T9 key was deactivated (without a dictionary), the proportion of textisms was .48. The children in the sample reported that they sent on average 24 SMSes per day and had been using a mobile phone for almost two years.

C. Bushnell, N. Kemp and F.H. Martin (2011) carried out a study with 227 Australian children between the ages of 10 and 12 (mean age: 11 years and 5 months). The children were asked to write down 30 conventional words just as they would if they were using the words in a message sent to a friend. Since the children were not allowed to use their telephone at school, the test was done with paper and pencil. The results indicated that the proportion of textisms in writing task was .53.

C. Wood, E. Jackson, L. Hart, B. Plester and L. Wilde (2011a) collected, in a longitudinal fashion over 10 weekends and half-term breaks, the SMSes of 56 children between the ages of 9 and 10 (mean age: 9 years and 10 months). The participants were from the Midlands region of the United Kingdom. The children had never used a mobile phone before the beginning of the study. Depending on the weekend, the number of messages sent varied between 6 and 45, and the proportion of textisms between .12 and .16. Over the 10 weeks, no evolution in usage was shown.

C. Wood, S. Meachmen, S. Bowyer, E. Jackson, M.L. Tarczynski-Bowles and B. Plester (2011b) carried out a longitudinal study of 1,019 children between the ages of 8 and 12 (mean age: 10 years and 4 months). The participants were from the West Midlands region of the United Kingdom. The children reported owning their mobile phone since the age of 8 years and 1 month. The children were asked to provide a sample of the messages they had sent at two different time periods: at the beginning of the school year and at the end of the school year. The results showed that the average ratio of textisms went from .33 to .40 between the beginning and the end of the school year. This slight increase masks the decreases at 8 to 9 years of age and at 11 to 12 years of age which remain unexplained. Both at the beginning and at the end of the year, the textism ratio was greater for the 11 to 12 year-olds than for the 8 to 9 year-olds (.42/.27 and .33/.074, respectively).

Overall, the studies carried out with children and young adolescents between the ages of 9 and 12 made use of a variety of methods ranging from the collection of natural data to the simulation of writing words in SMS language in a paper-and-pencil situation. The density of textisms revealed itself to also be highly variable: the values ranged from .074 to .53. Longitudinal studies of novice participants, such as those of Wood *et al.* (2011a), did not show an increase in the density of textisms with practice. This might be explained by the fact that a collection period limited to only 10 weekends was too short for an evolution to be demonstrated. It is therefore necessary to augment the existing data in order to have clear information regarding the orthographic forms of SMSes for the 9 to 12 year-old age group.

2. Research objectives

The existing research is limited and did not make use of the longitudinal method necessary for understanding the process of acquisition, with the exception of the studies carried out by C. Wood *et al.* (2011a) and C. Wood *et al.* (2011b). It should be noted that, in the first case, the study period was 10 weeks in length and, in the second case, the study period was one school year, but only one sample of SMSes is available for the beginning and the end of the year. To analyse the textisms, most

of the studies either used scenarios (N. Kemp and C. Bushnell 2011; B. Plester, C. Wood and P. Joshi 2009) or were not based on an SMS collection method in an ecological condition. The methodology used in the present study enabled the collection of SMSes in natural writing conditions. Furthermore, the categories of textisms could certainly be refined so as to be adapted to young texters. Finally, the existing studies did not take into account the gender variable (girls/boys) even though some research (J. Bernicot, O. Volckaert-Legrier, A. Goumi and A. Bert-Erboul 2012; A. Goumi, O. Volckaert-Legrier, A. Bert-Erboul and J. Bernicot 2011) show differences in SMS¹ length, dialogical structure and function in participants between the ages of 15 and 16. The objective of the present study was to fill in these gaps by examining, by means of a longitudinal study over a long period (12 months), how the orthographic characteristics of SMSes evolve (month by month) in young adolescents (11–12 years of age), by categorizing the textisms in terms of their distance from traditional written code.

3. Method

3.1 Participants

Nineteen young adolescents participated in the study: 10 girls and 9 boys (average age = 11.79 years, $SD^2 = .59$ of a year). They were recruited in a public secondary school in a town located in the Poitou-Charentes region of France (6th and 7th years). The pupils who had never owned or used a mobile phone were invited to participate in the study. The offer was as follows: to be equipped, free of charge, with a mobile phone for one year and to agree to “donate” to the research team at least 20 SMSes (written by the students themselves) per month. The research team guaranteed the students’ anonymity at all stages of the study. The students as well as their parents provided their written consent and agreement. The participants were all from middle class families, of legal school age, and native French speakers. The participants’ academic results in French class showed that they were able to write texts according to traditional rules, although they made some mistakes depending on their level.

1. The girls’ messages were longer than the boys’ messages and were often more relational than informational. The girls sent fewer messages which were lacking an opening and a closing (just the message) than the boys.

2. SD: Standard Deviation

3.2 Materials

The materials consisted of a Sony Ericsson model J132 or Alcatel model OT-303 mobile phones (French models). These two open-face (non-flip style) mobile phone models with a nine-key keyboard are similar. The T9 key, which enabled predictive text and access to a dictionary, was deactivated. The cards which served to recharge the mobile phones were, on the one hand, cards worth €15 (equivalent to 30 minutes of voice communication or 150 SMSes) and, on the other hand, cards worth €5 and valid for five days with unlimited SMSes. A 3G key and the Vodafone Mobile Connect software (a screenshot of this software is provided in Appendix A) installed on a computer allowed the research team to receive the SMSes which were ‘donated’ each month by the participants.

3.3 Procedure

The procedure for collecting the SMSes is summarized in Table 1. At the beginning of each month, the participants’ mobile phones were automatically credited with a sum of €15, the equivalent of 30 minutes of voice communication or 150 SMSes. Once per month, the participants’ mobile phones were also credited with the sum of €5, allowing an unlimited number of SMSes to be sent over a period of five days. It was during this period that the participants were to send at least 20 SMSes to the research team, freely chosen from the SMSes that they had sent throughout the month and that they had written themselves. These SMSes, with the help of the 3G key, were received onto a computer by means of the Vodafone SMS software. Using this software, a “cut and paste” procedure enabled the SMSes to be entered into the cells of a Microsoft Excel workbook. This procedure was repeated for 12 months.

Table 1. Stages of the SMS collection procedure

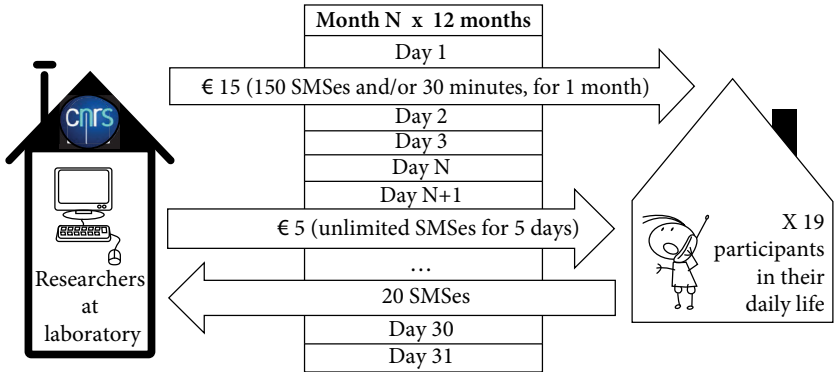


Table 2. SMSes from participant No 10 for month 1. The participant was a girl who was 12 years and 7 months old

Text message No	Original SMSes in French	English translations into traditional French of SMSes in transcribed form
1	Es ke jordan a un portable	<i>Does Jordan have a mobile phone</i>
2	Jlui et di je sai pa	<i>I told him I don't know</i>
3	Cc c megane Tu y sera ver kel heur au stade mercredi?	<i>Hey it's Mégane around what time will you be at the stadium Wednesday?</i>
4	3 ou 4	<i>3 or 4</i>
5	Chui ché des gen et je menui	<i>I'm at some people's house and I'm bored</i>
6	Ouai dmin et c ke laprèm Si tu vien fo ke tu révisé pars ky a un control en histoir	<i>Yeah t'morrow and it's that in the afternoon if you come you have to study because there's a history test</i>
7	Jai resu ton mésage mai pourquoi tu ma envoyé ça ?	<i>I got your message but why did you send me that?</i>
8	Cc J'ai une mauvaise nouvelle pour toi marine veu pa sortir avec toi Dsl	<i>Hey I have some bad news for you Marine does not want to go out with you sorry</i>
9	Ouai ta raison Tu me dira si ya des nouvel	<i>Yeah you're right tell me if you hear anything</i>
10	Non mai jvai à crepigny ché mon tonton et je vai mennuyé à mourir Pourquoi t au stade?	<i>No but I'm going to Crépigny to my uncle's place and I'm going to die of boredom Why are you at the stadium?</i>
11	A ok Ta fé té maths?	<i>Ah ok did you do your maths?</i>
12	Ladresse de ché ele?	<i>Her home address?</i>
13	Dsl jui pa a crepigny Jui en train de me baigné Mé normalemen jiré au foot vendredi prochain	<i>Sorry I'm not in Crépigny I'm swimming but I should be going to football next Friday</i>
14	A ok Et tu fé koi Ya un match	<i>Ah ok and what are you up to is there a match</i>
15	Nou on sor just de la piscine	<i>We just got out of the pool</i>
16	Cc dsl si je t pa répondu hier soir C parce ke jai étin mon portable	<i>Hey sorry I didn't get back to you last night it's because I turned off my mobile</i>
17	Es ke demain tu poura amené ta trouce ou ya tou t badge stp	<i>Tomorrow can you bring your case where you keep all your cards please</i>
18	Pk tu rép pa a ana	<i>Why won't you answer Ana</i>
19	Cc c vrai ke tu sor avec marine	<i>Hey is it true that you're going out with Marine</i>
20	Ba parce ke j aime bien tenbêté	<i>Ha because I really like messing with you</i>

The participants respected their agreement by providing a total of 4,524 messages:³ on average, per participant, 19.84 SMS ($SD=3.02$) per month and 238.10 SMS ($SD=36.24$) for the year. The data collection took place throughout the 2009–2010 academic year. As an example, Table 2 presents the SMSes (with their original form in French and their transcribed form in traditional French) of participant No 10 for month 1.

3.4 Coding

Two indexes were taken into consideration: the type of textisms and the density of textisms. A *textism* is defined as a change in the orthographic form of a word as compared to traditional writing. For each message, the *density* of textisms was equal to the number of words with changes divided by the total number of words in the message.

The coding of the textisms was based, on the one hand, on the analytical grids of the English language (R. Grinter and M.A. Eldridge 2003; B. Plester, C. Wood and P. Joshi 2009; C. Thurlow and A. Brown 2003) and on the grids dedicated to the French language (J. Anis 2007; R. Panckhurst 2009). R. Panckhurst (2010) showed the particularities of French (in SMS language) as compared to Italian and Spanish. E. Stark (2011) studied the morphosyntax in SMSes written in Swiss French, and examined SMSes in the three languages (French, German, Italian) spoken in Switzerland (E. Stark and C. Dürscheid 2011). In this study, it was deemed important to distinguish between the two large types of textisms with regard to their accordance with, or rupture from, traditional written code. From a cognitive point of view, the first case involves a different application of the same rules, while in the second case; there is an application or invention of different rules.

- a. The textisms which were consistent with the traditional code of grapheme–phoneme correspondence: the orthographic changes did not modify the phonology (pronunciation) of the words and were carried out with graphic forms which exist in traditional writing.
- b. The textisms which broke with the traditional code of grapheme–phoneme correspondence: the orthographic changes modified the phonology (pronunciation) of the words and/or were carried out with graphic forms which do not exist in traditional writing.

Table 3 presents the different sub-categories together with examples.

3. All of the text messages sent by the participants were taken into consideration, except for the text messages which were ‘chain letters’ or ‘spam’ (by definition not written by the participants themselves).

Table 3. Examples of different categories of textisms which were consistent with or breaking with the traditional grapheme–phoneme correspondence code

Textisms which were consistent with the traditional grapheme–phoneme correspondence code		
Categories	Examples	Traditional transcriptions in French
Simplifications	é	et
	donné	donner
	koi	quoi
	biz	bise
	tro	trop
	concer	concert
Complexifications	lessons	leçon
	j'ai male	j'ai mal
Substitutions	chanbre	chambre
	es	ai
	fais	fait
Textisms breaking with the traditional grapheme–phoneme correspondence code		
Categories	Examples	Traditional transcriptions in French
New grapheme–phoneme correspondences without modification of phonology	mwa	moi
	qi	qui
	dem1	demain
	ct	c'était
Agglutinations	entraïn	en train
	jcroyé	je croyais
Modifications of phonology	pk	pourquoi
	jtdgdt	je t'adore grave de trop
	couz	cousine
	chpa	je ne sais pas
Words or graphic forms which do not exist in traditional French	mdr	mort de rire
	jtkife	je t'aime
	ouiiiiiiiiiiiiiiiiiii	Repeated letters as an expression of urgency
	pfff	New onomatopoeia
	☹	Smiley expressing sadness

All of the categories of C. Thurlow and A. Brown (2003) were found in this study, except for “g-clippings”, which are specific to the English language. The category of “agglutinations”, made up of words placed one after the other without a space (“patavoir” [*nothaveyou*] instead of “pas t’avoir” [*not have you*], or “jcroyé” [*ithot*] instead of “je croyais” [*i thought*]), was highlighted, and this has no equivalent in the classification of C. Thurlow and A. Brown (2003).

With regard to the coding of the different types of textisms with the categories of this study, a very high intercoder agreement of 99.68% was found, using 250 messages chosen at random and containing 940 textisms.

4. Results

4.1 Different types of textisms

Table 4 provides an overall view of the distribution of the different types of textisms. In the corpus used in this study, the average proportion of textisms was .52, which meant that slightly more than half of the words produced by the participants contained a change with regard to traditional spelling and that slightly less than half contained no changes.

Table 4. Proportions of words with textisms for each category of textisms

Proportion of words with textisms	Proportion of words without textisms
.52	.48
Textisms which were consistent with the traditional code	
Categories	Proportions
Simplifications	.23
Complexifications	.01
Substitutions	.02
TOTAL	.26
Textisms breaking with the traditional code	
Categories	Proportions
New grapheme-phoneme correspondences without modification of phonology	.05
Agglutinations	.05
Modifications of phonology	.13
Words or graphic forms which do not exist in traditional French	.03
TOTAL	.26

Among the words with textisms, half (.26) were consistent with traditional code and the other half (.26) broke with this code. For the textisms which were consistent with the traditional code, simplifications made up the largest proportion (.23). For the textisms which broke with the traditional code, graphemes which led to a modification of the phonology made up the largest proportion (.13).

4.2 The density of textisms

First, with the participants as the random variable, the results were processed by means of a one-way ANOVA (analysis of variance): SMS experience (months 1 to 12). The dependent variable was the density of textisms: the total number of textisms divided by the number of words per message.

The effect of SMS experience was significant, $F(11,198)=3.90$, $p<.00004$, $\eta^2=.18$: on the whole, the density of textisms rose from month 1 (average = .44) to month 12 (average = .57) (cf. Figure 1).

For the textisms which were consistent with the traditional code, the results were processed by means of a two-factor ANOVA: SMS experience (months 1 to 12) x gender (female/male). The dependent variable was the density of textisms: the number of textisms consistent with the traditional code divided by the number of words per message.

The interaction effect between SMS experience and gender was significant, $F(11,187)=2.54$, $p<.005$, $\eta^2=.13$. For girls, this density (cf. Figure 2) remained stable over the months, while for boys, the increase between month 1 and month 12 was significant, $F(1,17)=6.00$, $p<.02$. Furthermore, for months 7 to 12 grouped together, boys produced significantly more textisms which were consistent with the traditional code than girls did $F(1,17)=6.40$, $p<.02$.

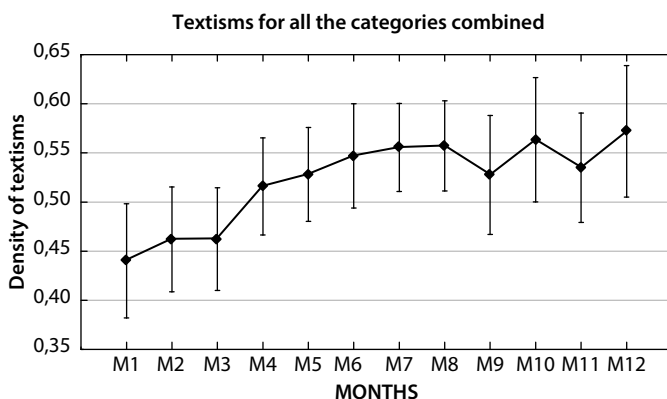


Figure 1. Evolution of textism density over one year by months (M) of SMS experience
The vertical bars represent the standard deviations

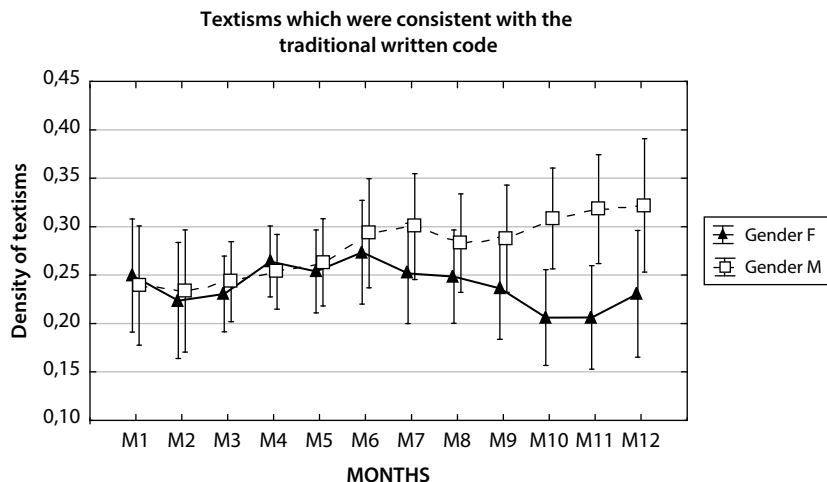


Figure 2. Evolution over one year of the density of textisms which were consistent with the traditional code by months (M) of SMS experience and gender (F: female, M: male) The vertical bars represent the standard deviations

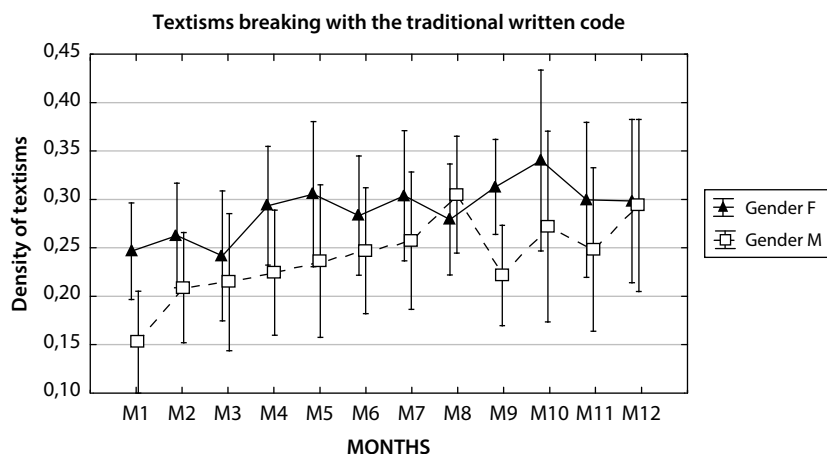


Figure 3. Evolution over one year of the density of textisms which broke with the traditional code by months (M) of SMS experience and gender (F: female, M: male) The vertical bars represent the standard deviations

For the textisms which broke with the traditional code, the results were also processed by means of a two-factor ANOVA: SMS experience (months 1 to 12) x gender (female/male). The dependent variable was the density of textisms: the number of textisms breaking with the traditional code divided by the number of words per message.

The effect of SMS experience was significant $F(11,187) = 2.40, p < .008, \eta^2 = .12$. No significant interaction effect for SMS experience $\times$ gender was found. For both girls and boys, there was an increase from month 1 to month 12 (cf. Figure 3).

Conclusions

The first objective of this study was to compile a corpus of SMSes written by young adolescents between the ages of 11 and 12. This goal was achieved: 4,524 SMSes produced by participants who had no experience at the start of the collection which lasted 12 months and included one collection per month. The technique which was used guaranteed that the SMSes were those written in daily life situations by the participants. The corpus, in French, added to that of C. Fairon, J.R. Klein and S. Paumier (2006) et E. Stark *et al.* (2011) ("sms4science") by providing data from younger texters (11 to 12 years of age), and whose productions were observed over a one-year period.

The first result was that the density of textisms increased with SMS experience (between month 1 and month 12 of the collection). In one year, the average grew from .44 to .57 with large variances: For example, in month 12, the smallest density value was .21 and the greatest value was .79. Differences, then, were observed from one participant to the other, but a general trend existed of a high density from the very first month and an increase in this density of textisms with SMS experience.

How might the high density of textisms produced by the novice users in the first month be explained? An initial explanation might point to a particularly weak level of spelling in traditional writing. Yet, as was mentioned earlier in this paper when the participants were presented, this was not the case: where spelling was concerned, all of these participants met the expectations of the academic system. A second explanation might be an immediate adaptation to the new writing context of the mobile phone: even if they had never used one themselves, the participants 'knew' by cultural transmission that SMS rules were not the same as classroom writing rules. This second hypothesis needs to be verified by means of a survey of novice users.

The average density of textisms that was observed (.57 after one year of experience) was greater than those found in comparable collections carried out with the longitudinal method: .16 for C. Wood *et al.* (2011a) and .40 for C. Wood *et al.* (2011b). Additional research will make it possible to determine whether this difference in results was due to the language (French/English) or the collection techniques. In the study carried out by C. Wood *et al.* (2011b), the density of textisms was .33 in the first month of the collection with participants who had, on average, two years of SMS experience; in the present study, it was .44 for participants

who had no experience at all. This study aimed to go further than an analysis of the density of textisms by considering the type of textisms and the gender of the participants. The distinction between the two types of textisms (consistent with or breaking with the traditional written code) was based upon the different cognitive processes which make a deeper analysis possible. On the whole, the density of textisms was .52, with a density of .26 for textisms which were consistent with the traditional written code, and the same density of .26 for those which broke with this code.

For the textisms which were consistent with the traditional code, the same rules were applied in a different way: for example, by simplifying “donner [*to give*]” and making it “donné”, the phoneme–grapheme code is respected “differently”. Instead of transcribing the “é” by the grapheme “er” (which is pronounced the same way) which is expected in terms of morphosyntax, it is transcribed by using another existing and phonologically possible grapheme, the letter “e”. The evolution with use is not the same for girls as for boys: for girls, the density remains stable, while for boys, it increases with use. For the last months (7 to 12), the density of textisms which were consistent with the traditional written code was higher for boys than for girls.

For textisms which broke with the traditional code, a new rule was applied or invented: For example, by writing “jtdgdt” for “je t’adore grave de trop” [*I love you way too much*], a message is produced which cannot be decrypted (or pronounced) using the traditional phoneme–grapheme traditional written code. A new rule is used, which leads to a new convention being used by those who have mastered the SMS register. The evolution with use is the same for girls as for boys: the density increases with use. At month 12, girls produced 57% of textisms which broke with the traditional code, while boys only made 47%. For this new variable, the results are analogous with the data obtained in other areas: girls tend to master the characteristics of an SMS register earlier than boys (J. Bernicot, O. Volckaert-Legrier, A. Goumi and A. Bert-Erboul 2012; A. Goumi, O. Volckaert-Legrier, A. Bert-Erboul and J. Bernicot 2011; C. Thurlow and M. Poff 2013).

This study demonstrated the importance of carrying out an in-depth analysis of textisms in terms of the cognitive processes involved by going further than the description of the orthographic changes with regard to the traditional written code. Continuing in this direction will make it possible to show and provide a convincing argument for the fact that SMSes are not an example of the deterioration of traditional writing, but instead constitute a register of their own which belongs to a very precise situation of communication (D. Crystal 2001; D. Ravid and L. Tolchinsky 2002): written correspondence with close friends and relatives by means of a mobile phone.

Acknowledgments

This research was supported by the French National Research Agency project ANR-08-COMM-011-01 entitled “Adolescents and Computer-Mediated Writing: Prerequisites, Uses, and Learning” and by the “Centre Henri-Aigueperse / UNSA Éducation” with the assistance of the IRES. Special thanks are extended to Martha Randy for translating this paper.

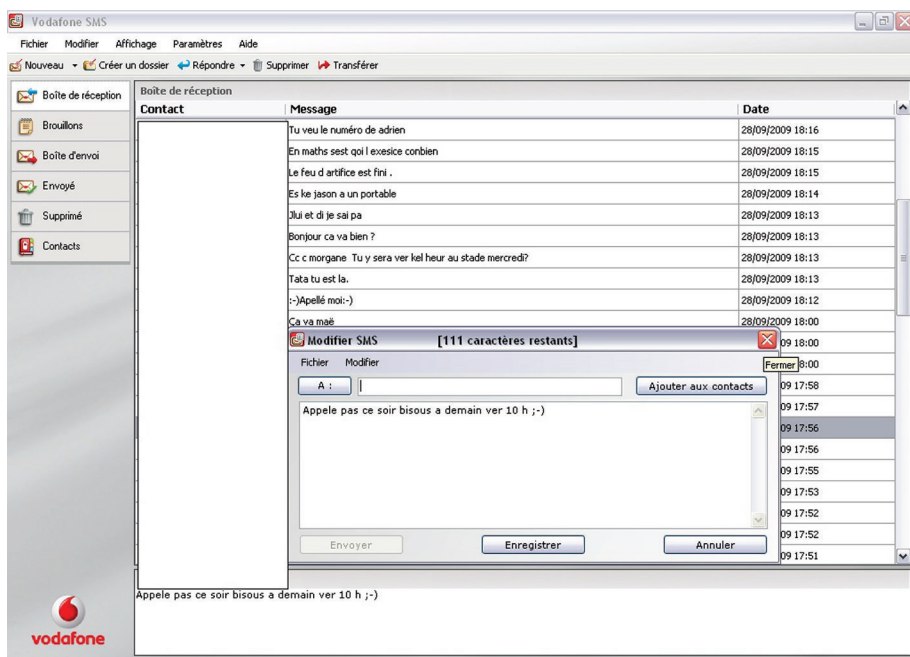
References

- Anis, Jacques. 2007. Neography: Unconventional Spelling in French SMS Text Messages. In *The Multilingual Internet: Language, Culture and Communication on Line*, B. Danet, S.C. Herring (eds.), New York: Oxford University Press.
- Austin, John L. 1962. *How to do Things with Words*. Cambridge, Massachusetts: Harvard University Press.
- Bernicot, Josie. 1994. Speech acts in young children: Vygotski's contribution. *European Journal of Psychology of Education* 9: 311–320.
- Bernicot, Josie; Alénouche Mahrokhian. 1989. To ask and to insist after a refusal: How do 6-7-year-old children proceed? *International Journal of Psychology* 3–4: 389–407.
- Bernicot, Josie; Judith Comeau; Helga Feider. 1994. Dialogues between French-speaking mothers and daughters in two cultures: France and Quebec. *Discourse Processes* 18: 19–34.
- Bernicot, Josie; Virginie Laval; Stéphanie Chaminaud. 2007. Nonliteral language forms in children: In what order are they acquired in pragmatics and metapragmatics? *Journal of Pragmatics* 39: 2115–2132.
- Bernicot, Josie; Olga Volckaert-Legrier; Antonine Goumi; Alain Bert-Erboul. 2012. Forms and functions of SMS messages: A study of variations in a corpus written by adolescents. *Journal of Pragmatics* 44: 1701–1715.
- Bigot, Régis; Patricia Croutte. 2011. *La diffusion des technologies de l'information et de la communication dans la société française*. Paris: CREDOC.
- Bushnell, Catherine; Nenagh Kemp; Frances Heritage Martin. 2011. Text-messaging practices and links to general spelling skill: A study of Australian children. *Australian Journal of Educational & Developmental Psychology* 11: 27–38.
- Crystal, David. 2001. *Language and Internet*. Cambridge: Cambridge University Press.
- Fairon, Cedrick; Jean René Klein; Sébastien Paumier. 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Goumi, Antonine; Olga Volckaert-Legrier; Alain Bert-Erboul; Josie Bernicot. 2011. SMS length and function: a comparative study of 13 to 18 year-old girls and boys. *European Review of Applied Psychology* 61(4): 175–184.
- Grice, Paul. 1975. Logic and Conversation. In *Syntax and Semantics 3: Speech Acts*, P. Cole, J. Morgan (eds.), New York: Academic Press.
- Grinter, Rebecca E; Margery A. Eldridge. 2001. Y do Tngrs Luv 2 Txt Msg? In *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work*, W. Prinz, M. Jarke, Y. Rogers, K. Schmidt, V. Wulf (eds.), Netherlands: Kluwer Academic Publishers.
- Kemp, Nenagh; Catherine Bushnell. 2011. Children's text messaging: abbreviations, input methods and links with literacy. *Journal of Computer Assisted Learning* 27: 18–27.

- Lenhart, Amanda; Ling Rich; Scott Campbell; Kristen Purcell. 2010. Teens and mobile phones. Pew internet and American life project. Retrieved 18 July 2012 from <http://www.pewinternet.org/~media/Files/Reports/2010/PIP-Teens-and-Mobile-2010-with-topline.pdf>
- Ling, Richard. 2005. The Socio-linguistics of SMS: An Analysis of SMS Use by a Random Sample of Norwegians. In *Mobile Communications: Renegotiation of the Social Sphere*, R. Ling, P.E. Pedersen (eds.), London: Springer.
- Noveck, Ira; Dan Sperber. 2004. *Experimental Pragmatics*. Basingstoke, England: Palgrave.
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, T. Arnavielle (ed.), Université Paul-Valéry: Montpellier 3.
- Panckhurst, Rachel. 2010. Texting in three European languages: does the linguistic typology differ?, *Actes du Colloque i-Mean 2009 Issues in Meaning in Interaction*, University of the West of England, Bristol, April.
- Plester, Beverly; Marja-Kristiina Lerkkanen; Lauri Linjama; Helena Rasku-Puttonen; Karen Littleton. 2011. Finnish and UK English pre-teen children's text message language and its relationship with their literacy skills. *Journal of Computer Assisted Learning* 27: 37–48.
- Plester Beverly A.; Clare Wood; Puja Joshi. 2009. Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *British Journal of Developmental Psychology* 27: 145–162.
- Ravid, Dorit; Liliana Tolchinsky. 2002. Developing linguistic literacy: a comprehensive model. *Journal of Child Language* 29: 419–448.
- Searle, John R. 1969. *Speech Acts*. Cambridge, MA: Cambridge University Press.
- Stark, Elisabeth, 2011. La morphosyntaxe dans les SMS suisses francophones: Le marquage de l'accord sujet — verbe conjugué. *Linguistik Online* 48(4). Retrieved 18 July 2012 from http://www.linguistik-online.de/48_11/stark.htm.
- Stark, Elisabeth; Christa Dürscheid, 2011. SMS4science: An International Corpus-based Texting Project and the Specific Challenges for Multilingual Switzerland. In *Digital Discourse. Language in the New Media*, C. Thurlow, K. Mroczek (eds.), 299–320, Oxford: Oxford University Press.
- Tossell, Chad C.; Philip Kortum; Clayton Shepard; Laura H. Barg-Walkow; Ahmad Rahmati; Lin Zhong. 2011. A longitudinal study of emoticon use in text messaging from smart-phones. *Computers in Human Behavior* 28: 659–663.
- Thurlow, Crispin; Alex Brown. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. Retrieved 28 July 2010 from [http://faculty.washington.edu/thurlow/papers/Thurlow \(2003\)-DAOL.pdf](http://faculty.washington.edu/thurlow/papers/Thurlow%20(2003)-DAOL.pdf)
- Thurlow, Crispin; Michele Poff. 2013. The Language of Text Messaging. In *Handbook of the Pragmatics of CMC*, S.C Herring, D. Stein, T. Virtanen (eds.), 163–189. Berlin and New York: Mouton de Gruyter.
- Verschueren, Jeff. 1999. *Understanding Pragmatics*. London: Arnold.
- Wood, Clare; Emma Jackson; L Hart; Beverly Plester; Lucy Wilde. 2011a. The effect of text messaging on 9- and 10-year-old children's reading, spelling and phonological processing skills. *Journal of Computer Assisted Learning* 27: 28–36.
- Wood, Clare; Sally Meachem; Samantha Bowyer; Emma Jackson; M. Luisa Tarczynski-Bowles; Beverly Plester. 2011b. A longitudinal study of children's text messaging and literacy development. *British Journal of Psychology* 102(3): 431–442.

Appendix

Screenshot of the software enabling the researchers to receive the SMSes sent by the participants



Summary

The aim of this paper was to study the characteristics of SMSes in a population for which there is currently only limited data: young adolescents (girls and boys) between 11 and 12 years of age. The analysis focused on a corpus of 4,524 SMSes sent by 19 informants in everyday real-life situations over a one-year period. At the beginning of the study, the participants were complete novices. This study sets forth a new analysis grid which distinguishes between two categories of textisms which were defined based on the following cognitive processes: (a) textisms which are consistent with the traditional written code of grapheme–phoneme correspondence and (b) textisms which break with this traditional code. On the whole, the density of textisms was .52 and .26, respectively, for each kind of textism. The results showed an increase in the density of textisms with SMS experience (from month 1 to month 12), but also a variation depending upon the type of textism and the gender of the texter. For boys, the density of both types of textisms increased with SMS experience, while for girls, the density of textisms only increased for textisms which broke with the code. The results were interpreted in terms of the construction of an SMS register with specific linguistic markers resulting from a different use of traditional writing rules or the use of inventions as compared to traditional writing.

Automatic or Controlled Writing?

The Effect of a Dual Task on SMS Writing in Novice and Expert Adolescents

Céline Combes, Olga Volckaert-Legrier and Pierre Largy
Université Toulouse II — Le Mirail Laboratoire Psychologie du
Développement et Processus de Socialisation — EA 1687

Introduction

Writing has been the focus of study for a number of years (C. Bereiter and M. Scardamalia 1987; J.R. Hayes and L. Flower 1980), essentially through cognitive and psycholinguistic approaches. The activity of writing has been modeled, both in experienced adults (J.R. Hayes 1996; R.T. Kellogg 1996) and in children who are still learning (V.W. Berninger and H.L. Swanson 1994; D. McCutchen 1996). These models have made it possible to identify the various cognitive processes which come into play in the writing of a text. The simultaneous implementation of these different processes (*i.e.*, *planning, formulation, revision*) is a complex, cognitively demanding activity, especially in children who are learning to write (B. Bourdin 1999). The difficulty of this activity stems from the writer's limited ability to process the working memory-associated information required to produce a written piece of work (M.J. Power 1985). These writing processes were first studied in terms of the production of a text. For several years now, the focus of research on writing processes has been on the analysis of smaller linguistic units (*i.e.*, *sentences*), leading then to studies of orthographic production (M. Fayol, P. Largy and P. Lemaire 1994; P. Largy, M.-P. Cousin and A. Dédéyan 2005). However, all of the research has been focused on the most conventional form of writing, in other words, the written manuscript.

With the arrival of such modes of communication as e-mail or the SMS (short message service), different communication tools are now being used for writing (computer keyboard, alphanumeric keyboard). The SMS is one new kind of writing which is carried out despite various constraints, such as a limited number of characters and an alphanumeric keyboard (J. Anis 2001, 2002; C. Fairon, J.R. Klein and S. Paumier 2006a). As a result, new orthographic forms have emerged and

have been categorized (J. Anis 2002; D. Crystal 2001; C. Fairon, J.R. Klein and S. Paumier 2006b; F. Liénard 2008; R. Panckhurst 2009). The study of the nature of these new orthographic forms suggests that different processes come into play (*i.e.*, *phonological*, *word retrieval*) as the forms are being produced.

This study attempts to distinguish between the different processes involved in producing SMS orthographic forms. The production of these different orthographic forms was compared using an experimental paradigm: the dual task. This paradigm requires that two tasks are implemented at the same time. This is intended to force the attentional resources required for the orthographic production of SMSes, thereby enabling an understanding of the controlled and/or automatic processes related to writing. The degree of automation of these processes in the production of SMSes was also examined with regard to the texter's level of SMS writing expertise.

1. Writing and working memory

While the control and automation of cognitive processes in writing have barely been studied in terms of SMS writing (eSMS¹), they have been studied with regard to conventional writing (M. Fayol and A. Miret 2005; R.T. Kellogg 1994). Early studies of conventional writing studies demonstrated that carrying out a writing activity entailed functional constraints. The main constraint was a writer's limited ability to process information associated with working memory (C. Bereiter and M. Scardamalia 1987; L.S. Flower and J.R. Hayes 1980). Writing involves controlled processes (*i.e.*, *planning and revision*) which take place consciously and necessitate a cognitive effort as well as automatic processes (*i.e.*, *formulation process as orthographic production*) which call upon fewer attentional resources (D. Alamargot, E. Lambert and L. Chanquoy 2005; W. Schneider and R.M. Shiffrin 1977). Working memory (WM) is thought to be implicated in the temporary maintenance of processed information. For example, up until the moment of its material realization, the texter must temporarily store the message. Research by A.D. Baddeley (1986;1992) defined the three components of WM: a central administrator aided by two auxiliary systems, the phonological loop and the visuo-spatial sketchpad. The central administrator is a limited system which is responsible for allocating sufficient attentional resources for the completion of a task. Furthermore, this system may be mobilized for cognitively advanced tasks (*i.e.*, *language comprehension, etc.*). The phonological loop is responsible for the maintenance of verbal

1. The abbreviation for the term "SMS-type writing", "eSMS", was first used by R. Panckhurst (2009) and is used throughout this article.

information, and consists of a phonological storehouse and an articulatory process. The phonological storehouse immediately receives verbal information which has been heard while the articulatory process allows for information to be updated by re-entering the storehouse. Finally, the visuo-spatial sketchpad is used for the storage and temporary maintenance of visual and spatial information.

In the early stages of learning, some specific writing activities (*i.e.*, *spelling, graphic symbols, etc.*) have not yet become automatic. The participation of WM, therefore, is very high where these activities are concerned (B. Bourdin 1999). Indeed, the processes involved in writing require conscious control until they become automatic, thereby mobilizing attentional resources and interfering with the implementation of more complex controlled processes (W. Schneider and R. M. Shiffrin 1977). Writing on a regular basis enables the automation of the spelling process and lightens the cognitive cost linked to this activity. This cost is especially great for novices, and decreases as expertise is acquired, thereby freeing up WM resources which can be allocated to more complex tasks (R.T. Kellogg 1987). Some orthographic processes, such as subject-verb agreement, become automatic as the writer acquires expertise, thus allowing the writer to deal with the agreement using fewer and fewer cognitive resources. In contrast, other processes prove more difficult to render automatic, such as complex grammatical agreement (*i.e.*, *agreement of the past participle with the auxiliary verb "avoir" [to have]*) the implementation of which remains cognitively demanding for quite some time (C. Brissaud and J.-P. Chevrot 2001).

Controlled processes in writing are primarily understood by means of the dual-task paradigm (M.J. Power 1985). This paradigm, requiring that two tasks be carried out simultaneously, enables the important role of WM in the production of writing to be demonstrated. For example, M. Fayol, P. Largy and P. Lemaire (1994) asked pupils to carry out two tasks in parallel: to write out one dictated sentence (primary task) and to repeat the words from a list which they had already heard (secondary task). The secondary task mobilized one part of the participants' attention at the very moment when they were to carry out the primary task, thereby preventing them from focusing all their attention on the primary task. This paradigm creates a conflict in the allocation of attentional resources since the two tasks mobilize the same kind of resources. The use of this paradigm makes it possible to measure the cognitive cost of producing a written text (M. Fayol and P. Largy 1992; M.J. Power 1985).

Studies which have analyzed the cognitive cost of producing writing have focused on conventional writing. With the arrival of new communication tools, writing has evolved, leading to the emergence of new orthographic forms. Research into these new written forms implies that different processes come into play as they are being produced, and this research examines these different processes of orthographic SMS production.

2. SMS and spelling

Few studies have shown interest in the relationship between SMS and spelling, and most research has been carried out on English-speaking children. In a study by B. Plester, C. Wood and V. Bell (2008), 35 pupils between the ages of 10 and 11 were asked to transcribe an eSMS text into conventional English, and vice versa. The results showed a positive relationship between eSMS language and spelling. Indeed, when pupils scored well in conventional spelling (measured by the British Ability Scales II — C.D. Elliot, P. Smith and K. McCulloch 1996), they also performed well in eSMS (few interpretation errors and a good ratio of ‘textisms’). According to B. Plester *et al.* (2008), a solid knowledge of standard writing codes and rules is necessary in order to be able to “play with words” and thereby produce orthographic modifications such as in SMS.

In a follow-up study, C. Wood *et al.* (2011) recruited 114 pupils between the ages of 9 and 10 who had never used a mobile phone and assigned them to either an experimental or a control group. The children in the experimental group were provided with mobile phones so that they could send SMSes over the week-end and after school hours for a period of 10 weeks. The children in the control group were also given access to the phones for a brief period, so that they could also experience using them. The results obtained did not show a difference in the level of spelling between the two groups of participants at the end of the study. For the experimental group, a positive correlation was observed between the number of eSMS modifications and an improvement in their reading and writing abilities. The results showed that eSMS did not negatively impact the development of the reading and writing abilities of children between the ages of 9 and 10. The results also showed the use of eSMS is positively linked to improved writing ability, especially in spelling.

A search of published literature revealed that only one other study made use of the dual-task paradigm with regard to SMS language writing (C. Combes, O. Volckaert-Legrier and P. Largy 2012). In this study, 24 French-speaking pupils in the 6th year were asked to write out SMSes that were dictated to them in two different situations. In the single-task situation, the pupils listened to a recorded message and were asked to write it on their cell phone. In the dual-task situation, they heard audio beeps while they were writing the message. After having written the message, they were asked to recall the total number of audio beeps they had heard. Although they did not demonstrate the dual-task effect, the results obtained showed that the expert texters produced more eSMS-specific modifications (*i.e.*, *reductions*) than the novice texters. Furthermore, the novice texters made more ‘absence’-style modifications at the end of the message than the experts. C. Combes *et al.* (2012) concluded that eSMS becomes more automatic as the users acquire expertise. The goal of the beep-listening task used in this study was to

mobilize the attentional resources of the phonological loop. However, the cognitive cost borne by participants when performing this task was not great enough to disrupt the phonological loop. A different dual-task study (M. Fayol, P. Lamy and P. Lemaire 1994) showed that a secondary task consisting of pseudowords presented a cognitive cost which was great enough to interfere with the phonological loop. Thus, a secondary task such as this one was chosen for this study.

Objectives

The existing research (C. Combes *et al.* 2012; B. Plester *et al.* 2008; C. Wood *et al.* 2011) shows that eSMS expertise is acquired through the frequent use of this new mode of communication. Furthermore, the use of this new form of writing has been seen to improve spelling knowledge and expertise in pre-teens. As we stated earlier, however, the effect of the secondary task was not demonstrated in the research carried out by C. Combes *et al.* (2012) in which an attempt was made to study the cognitive cost which this writing represents. This is why we adapted their experimental protocol in an attempt to distinguish the different processes involved in producing SMS orthographic forms. We compared the production and the potential automation of these different orthographic forms by means of an experimental dual task paradigm and the comparison of SMS production with regard to eSMS expertise.

3. Method

After describing the participants' characteristics, the measures used during the two sessions of this study will be presented. Finally, the experimental procedure will be exposed.

3.1 Participants

The participants were 19 children enrolled in the 6th year (median age: 11.5 years, *S.D.*: 3.57). To ensure that the modifications observed in the SMSes were not due to a weak spelling level, only participants who had not repeated a year of school and who had the prerequisite level of spelling (measured using the Spelling Level Test — TNO) were chosen. Based upon their answers, participants were assigned to a novice group, which sent an average of 20 SMSes per week and had owned a mobile phone for at least one year, and an expert group, which sent between 50 and 100 SMSes per week and had owned a mobile phone for more than two years.

3.2 Measures

In the first part of this experiment, we used the secondary Spelling Level Test (TNO — *Test de Niveau d'Orthographe* — F. Doutriaux and R. Lepez 1980) and a questionnaire measuring the participants' typical SMS use. Then, upon implementing the experimental protocol, we used the SMS messages for the primary task and the pseudowords for the secondary task.

3.2.1 Spelling Level Test (TNO)

To measure the participants' spelling level, we used the secondary form of the TNO for 6th year pupils. This test was made up of 90 items separated into two kinds of exercises. The first kind of exercise (45 items) involved identifying the correctly spelled word from three choices (1). In the second kind of exercise (45 items), the pupil was asked to identify the incorrectly spelled word from three choices (2a–c). Each of these exercises was composed of 50% usage-based spelling² items and 50% grammar-based spelling³ items. The secondary form of this test was standardized for classes from the 6th year until the final year of secondary school.

- (1) *Je sens une _____ qui me grimpe dans le dos .* fourmis – fourmie – fourmi
(I feel that an _____ is climbing on my back) ants – ant (The second word is a phonetic distraction)
- (2) a. *Il portait de magnifiques souliers marrons.*
(He wore beautiful brown shoes.)
- b. *Les pantalons verts lui plaisaient beaucoup.*
(He liked green trousers very much.)
- c. *Ce bébé aura sûrement les yeux noisette.*
(This baby will certainly have hazel eyes.)

3.2.2 Questionnaire

The questionnaire enabled the creation of novice and expert groups based on the participants' length of experience sending SMSes and the average number of SMSes they sent per week. The questionnaire was an adaptation of the one used in the "SMS4Science" project (C. Fairon *et al.* 2006b) and provided complementary information regarding the participants' understanding of SMS language as well as their user habits with other modes of communication (e-mail, instant messaging, chat, etc.). The questionnaire was composed of 10 closed questions (five dichotomous and five multichotomous with only one answer) as well as a section

2. Spelling which can be checked in a dictionary.

3. Spelling which results from the implementation of grammar rules.

about the individual characteristics of the participants (age, gender, year in school, whether they had repeated a school year).

3.2.3 *Materials for the experimental protocol*

The materials used for the experimental protocol included the SMS messages for the primary task as well as a list of pseudowords for the secondary task.

a. Messages

Thirty-two messages were created (2a–d). In a classic communication situation, an interaction takes place in three steps: the opening, the body and the closing (V. Traverso 1999). In this vein, each message produced for this study contained an opening, a message body and a closing (1). A pronoun, a verb and a noun were present in the body of each message. All the messages which were created for this study had less than 160 characters (the maximum size of a message) so as to leave the participants free to make spelling modifications (or not) depending upon their SMS writing habits.

- (1) *Opening / Message body / Closing*
- (2) a. *Coucou / Pourquoi tu m'envoies ce message? / Bisous*
(Hey / Why are you sending me this message? / Kisses)
- b. *Salut / C'est bon je vois ta voiture / A toute*
(Hi / No problem I can see your car / Later)
- c. *Bonjour / J'ai enfin rechargé mon crédit / A plus*
(Hello / I finally recharged my credit / See you)
- d. *Bonsoir / Qu'est-ce que tu dis de cette idée? / Bises*
(Evening / What do you think of this idea? / Kisses)

The 32 messages were recorded by a speaker from Toulouse, France, because of the local accent,⁴ so as to guarantee that all the participants were presented with the identical material.

b. Secondary task

Sixteen lists of three monosyllabic pseudowords were created for the secondary task (3a). Each pseudoword was composed of 3–4 phonemes. A list of three pseudowords was associated with a message in the dual task situation (3b–c).

- (3) a. poic — trit — sonc
- b. Message / Secondary task (three pseudowords)
- c. *Coucou. Pourquoi tu m'envoies ce message ? Bisous / poic – trit – sonc*
(Hey / Why are you sending me this message? / Kisses / poic – trit – sonc)

4. The experiment was conducted in the Toulouse area.

4. Procedure

First, at the initial session, the children were tested as a group by completing the questionnaire and the TNO. At the second session, the children participated individually by means of the experimental protocol. Finally, the data stemming from this second session were coded. These different steps are explained in detail below.

4.1 Questionnaire and TNO

The researcher reads each question to the participants so as to be sure that the participants had understood.

The administration of the TNO was split into two sections in order to ensure that the participants did not confuse the instructions about the two types of exercises. No time limit was imposed upon the participants for the test. The only results which were kept were those of the children with a good spelling level, *i.e.*, a score of 27 or higher on the TNO for 6th year pupils.

4.2 Experimental protocol

During the administration of the experimental protocol, the participants were confronted with two experiments: one single-task test and one dual-task test. The order in which the two conditions were presented alternated from one participant to the next. The 32 messages were divided into two sets (A and B) of 16 messages each. Half of the participants were presented with set A for the single task and set B for the dual task and vice versa. Four groups of participants were created according to the order in which the sets were presented and the experimental situation (single task vs. dual task). Each of these four groups was made up of 50% novice and 50% expert eSMS users.

The participants carried out the experimental protocol on their own mobile phones. This condition was chosen so as to replicate as closely as possible their natural conditions of SMS use and so as to avoid an additional cognitive cost linked to the use of new material.

In the single task, the participants listened to a message. After listening to the message, they were asked to write it on their mobile phone as if they were sending an SMS to someone. The message could only be heard once. In the dual-task condition, the participants listened to a list of three pseudowords after the message. After writing the message on their cell phone, the participants were asked to recall and write down in a notebook the three pseudowords they had heard. A training phase was offered to the participants.

After each message was produced, the cell phone screen was photographed.

4.3 Coding

The spelling modifications produced in the messages were analyzed and the spelling modifications corresponded to the variations from normal written spelling. For each participant and for each message, the number of spelling modifications was tallied and related back to the number of words present in the experimental message. Since the number of words contained in each message varied, the number of spelling modifications which were observed in each message was compared with the message's total number of words. The results were presented as percentages.

These spelling modifications were categorized using a classification based upon the typology by R. Panckhurst (2009), and are presented in Table 1.

Table 1. Classification adapted from the typology by R. Panckhurst (2009)

Categories	Examples
I. Substitutions, simplifications, deletions	
– Complete	“c” (“c’est”) [it is], “g” (j’ai) [I have]
– Partial	”ossi” (aussi) [also], “bizes” (bises) [kisses]
– Diacritic symbols	“lecons” (leçons) [lessons], “à” (à) [to]
– Elision	“est ce que” (est-ce que) [is it], “j ai” (j’ai) [I have]
– Icons, mathematic symbols, special characters...	à + (à plus), @ [see you later]
II. Reductions	
– Morpho-lexical abridgment (truncations, abbreviations/acronyms)	“exos” (exercices) [exercises], “mess” (message) [message]
– Reduction of the grapheme “qu”	“ke” (que) [that], “pourkoi” (pourquoi) [why]
– Elimination of the silent endings of words	“bis” (bise) [kiss]
– Consonant skeletons	“cc” (coucou) [hi], “slt” (salut) [hi]
– Agglutinations	“jai” (j’ai) [I have], “jvois” (je vois) [I see]
III. Lexical and grammatical spelling	
– Verbal ending	“tu parle” (tu parles) [you say] + inf / pp confusion
– Nominal ending	“tes dessin” (tes dessins) [your drawings]
– Spelling error	“experience” (expérience) [experience]
V. Other modifications	
– Mechanical errors (typos)	“ayitude ?” (attitude) [attitude]
– Additional characters: emoticons, smileys	“oki” (ok) [ok], “les zexercices” (les exercices) [the exercises]
– Replacement with another formulation	“A plus” (A toute) [see you later]
– Absence	

In order to validate the typology based upon the adaptation of R. Panckurst's (2009) typology, the data were double coded. Intercooder agreement was carried out on 10% of the corpus by two independent judges. The intercooder coefficient was evaluated using the Bravais-Pearson correlation coefficient ($r=0.95$).

5. Results

For each dependent variable, an analysis of variance (ANOVA) was used to carry out the statistical analyses. The results took into account the percentage of spelling modifications per complete message, as well as the type of spelling modifications made. The types of spelling modifications were studied in an effort to distinguish the different processes by which the SMS spelling forms were produced. The spelling modification categories were used to examine of the types of spelling modifications. Only categories with more than 5% of spelling modifications were studied.

5.1 Analysis of the percentage and type of modifications

The total percentage and the type of spelling modifications were analyzed by means of a four-factor ANOVA: Expertise (Novice; Expert), Message section (Openings; Message bodies; Closings), Task (Simple task; Dual task), Modification category (Substitutions; Reductions; Verbal and nominal endings; Other modifications) with repeated measures for the last three factors. The percentages were calculated based upon the number of spelling modifications per number of words in the observed part of the message. Since some words contained several spelling modifications, the results obtained were sometimes greater than 100.

The analysis showed a significant effect of expertise: $F(1, 17)=6.95$, $p<.05$, $\eta^2_p=.29$. The expert texters produced more spelling modifications than the novice texters (74.1% vs. 57.6%).

The effect of the message section was also significant: $F(2, 34)=12.03$, $p<.001$, $\eta^2_p=.41$. On average, the participants made more spelling modifications in the closing than in the body of the message: $F(1, 17)=5.43$, $p<.05$ (respectively: 88.2% vs. 67.7%). They made more spelling modifications in the closing than in the opening: $F(1, 17)=15.68$, $p<.01$ (88.2% vs. 50.1%). They made more spelling modifications in the body of the message than in the opening: $F(1, 17)=9.50$, $p<.01$.

The effect of the task was significant: $F(1, 17)=7.70$, $p<.05$, $\eta^2_p=.31$. The participants produced more spelling modifications in the dual-task situation than in the single-task situation (69.7% vs. 67.7%).

The effect of the spelling modification categories was significant: $F(3, 51)=12.34$, $p<.001$, $\eta^2_p=.42$. The participants made more "other" modifications

(142.4%) and reduction-type modifications (127.9%) than substitutions (92.1%) and modifications of verbal and nominal endings (37.7%). All four simple effects are presented in Figure 1.

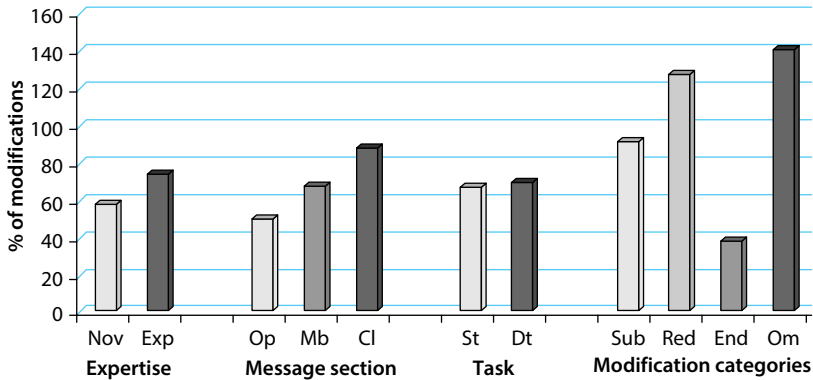


Figure 1. Percentages of modifications by expertise (Nov: Novice, Exp: Expert), message section (Op: Opening, Mb: Message body, Cl: Closing), task (St: Single task, Dt: Dual task), category (Sub: Substitutions, Red: Reductions, End: Endings, Om: Other modifications)

The analysis showed a significant interaction between the message section and the modification categories: $F(6, 102) = 30.41$, $p < .001$, $\eta_p^2 = .64$. In the opening, the participants made more reductions (35.2%) than other modifications (13.2%): $F(1, 17) = 6.95$, $p < .05$. However, they did not use substitutions or modifications of verbal and nominal endings. In the message body, the participants primarily made substitutions (20.1%), reductions (19.1%), verbal and nominal ending modifications (18.7%), as well as other kinds of modifications (13.5%): $F(3, 51) = 3.45$, $p < .05$. In the closing, the most frequently used modifications were other modifications (44.4%), substitutions (25.9%), and reductions (9.7%): $F(2, 34) = 17.86$, $p < .001$. In the closing, no verbal and nominal ending modifications were made.

The interaction between the task and the modification categories was significant: $F(3, 51) = 18.58$, $p < .001$, $\eta_p^2 = .52$. The participants made significantly more substitutions while doing the single task than doing the dual task (26.4% vs. 19.6%): $F(1, 17) = 7.75$, $p < .05$. In contrast, they made a greater number of other modifications while doing the dual task than the single task (30.6% vs. 16.9%): $F(1, 17) = 23.47$, $p < .001$. The difference in production of reductions and verbal and nominal ending modifications between the single task and the dual task was not significant: $F_s < 1$, ns.

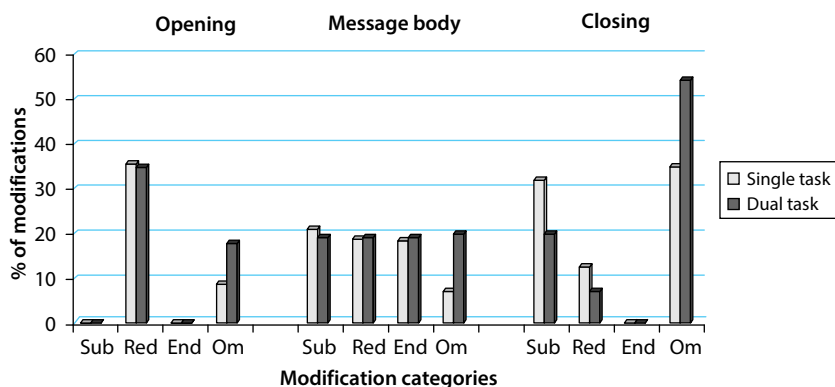


Figure 2. Percentages of modifications by category (Sub: Substitutions, Red: Reductions, End: Endings, Om: Other modifications) according to task type and message section.

The three-way interaction between the message section, the task and the modification categories was significant: $F(6, 102) = 3.12$, $p < .01$, $\eta^2_p = .16$. This effect is presented in Figure 2. In the openings, the differences between the single task and the dual task were not significant for reductions (35.5% vs. 34.9%): $F < 1$, ns and other modifications (8.9% vs. 17.7%): $F(1, 17) = 3.38$, ns. In the message bodies, the participants made more modifications in the dual-task situation (19.8%) than in the single-task situation (7.2%): $F(1, 17) = 15.57$, $p < .01$. The differences between the single task and the dual task were not significant for substitutions (20.9% vs. 19.3%), reductions (18.9% vs. 19.3%) and modifications of verbal and nominal endings (18.4% vs. 18.9%) in the message body: $F_s < 1$, ns. In the closing, the participants made more substitutions in the dual-task situation: $F(1, 17) = 9.53$, $p < .01$ (31.9% vs. 19.9%). They also made more reductions in the single-task situation than in the dual-task situation: $F(1, 17) = 11.99$, $p < .001$ (12.2% vs. 7.2%). In contrast, they made a greater number of other modifications in the dual-task situation than in the single-task situation: $F(1, 17) = 15.71$, $p < .01$ (54.2% vs. 34.7%).

No other significant interaction effect was found, $F_s < 1$, ns.

Since the categories themselves were made up of sub-categories (see Table 1), the variance analysis was continued by replacing the four-mode categories factor by the 17-mode sub-category factor. The fact that these sub-categories were not homogeneous created the expectation that the spelling processes related to their production would be different.

5.2 Analysis of the spelling modification sub-categories

With regard to the spelling modification sub-categories, a four-factor ANOVA was carried out: Expertise (Novice; Expert), Message section (Opening; Message

body; Closing), Task (Simple task; Dual task), Sub-categories (17) with repeated measures for the last three factors. The dependent variable was the percentage of modifications per sub-category type.

The simple effects of expertise, the task and the message sections were the same as in the preceding analysis.

The effect of the modification sub-categories was significant: $F(16, 272) = 14.50$, $p < .001$, $\eta_p^2 = .46$. The sub-categories of substitution which the participants made the most were icons (15.1%). The reductions which were made the most were consonant skeletons (12.8%). The participants often made verbal-ending modifications (11.4%). Where the other modifications were concerned, the most commonly used sub-categories were absences (12.1%) and replacements by another formulation (9.1%). The other modification sub-categories were not presented as the participants made these modifications less than 5% of the time.

The interaction between the message section and the modification sub-categories was significant: $F(32, 544) = 18.05$, $p < .001$, $\eta_p^2 = .52$. In the opening, the participants typically made consonant skeletons (29.2%). The other sub-categories were of spelling modifications which the participants made less than 5% of the time. In the message body, the sub-categories of the most commonly used modifications were verbal endings (11.4%), absences (6.7%), and consonant skeletons (6.4%). Finally, in the closing, the modifications which were made the most were absences (28.2%), icons (24.1%) and replacements by another formulation (15.6%).

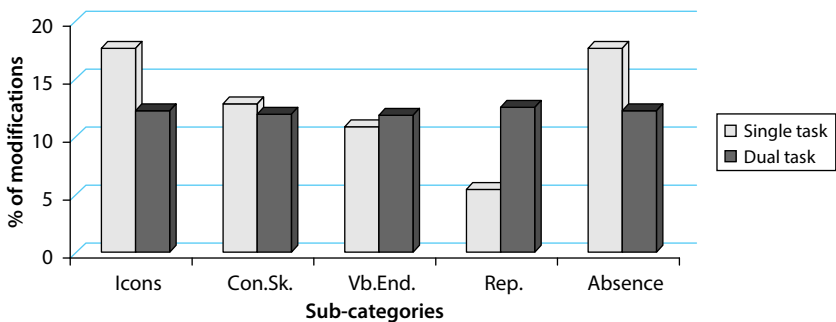


Figure 3. Percentages of modifications by sub-categories (Icons, Con. Sk.: Consonant Skeletons, Vb. End.: Verb Endings, Rep.: Replacement by another formulation, Absence) according to task type

The interaction between the task and the modification sub-categories was significant: $F(16, 272) = 11.28$, $p < .001$, $\eta_p^2 = .40$. The participants made more icons in the single-task situation (17.8%) than in the dual-task situation (12.4%): $F(1, 17) = 7.14$, $p < .05$. In contrast, the participants made more replacements with

another formulation ($F(1, 17) = 40.70, p < .001$; 5.5% vs. 12.7%) and absences ($F(1, 17) = 13.21, p < .01$; 8.2% vs. 15.8%) in the dual-task situation than in the single-task situation. The differences between the single task and the dual task were not significant for consonant skeletons (13.1% vs. 12.1%) or modifications of verbal endings, $F_s < 1$, ns. Figure 3 presents these results.

The interaction between the message section, the modification sub-categories and expertise was significant: $F(32, 544) = 1.77, p < .01, \eta_p^2 = .09$. In the opening, the expert texters produced only slightly more consonant skeletons than the novice texters (42.7% vs. 15%: $F(1, 17) = 3.56, p = .07$. No other effect was significant: $F_s < 1$, ns.

The interactive effect between the message section, the task and the modification sub-categories was significant: $F(32, 544) = 2.93, p < .001, \eta_p^2 = .15$. In the opening, no significant interaction effect between task and consonant skeletons was found, $F < 1$, ns (29.3% vs. 27.6%). In the message body, the participants produced significantly more absences in the dual-task situation (10.6%) than in the single-task situation (2.9%): $F(1, 17) = 16.54, p < .001$. The same effect was observed in the closing: $F(1, 17) = 10.60, p < .01$ (35.2% vs. 21.2%). In the closing, the participants produced more icons in the single-task situation than in the dual-task situation (29.2% vs. 18.9%): $F(1, 17) = 7.11, p < .05$. Figure 4 presents these results. No other effect was found to be significant: $F_s < 1$, ns.

No other interactive effect was found to be significant: $F_s < 1$, ns.

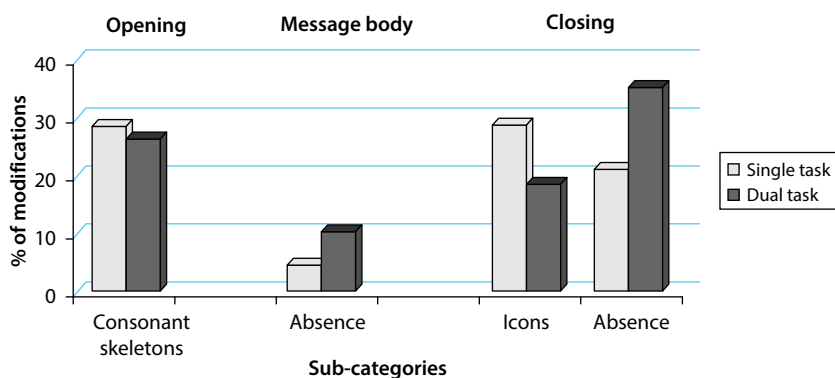


Figure 4. Percentages of modifications by sub-category according to task type and message section

Discussion

In this study, the objective was to compare the production of different eSMS spelling forms and their possible automation. To accomplish this, the degree of SMS writing expertise of the participants and the attention they were able to devote to SMS writing through the experimental paradigm of the dual task were varied.

First of all, a possible automation of the SMS production processes was noted as eSMS expertise was acquired. Indeed, the results obtained demonstrated that the eSMS experts had a tendency to produce more “consonant skeleton” modifications in openings than the novice texters. This type of modification resembles that of note-taking (F. Mourlhon-Dallies, 2010); in the 6th year, however, pupils are not yet used to taking notes. These modifications therefore appear specific to eSMS. These results concur with those of B. Plester *et al.* (2008). In their study, the participants primarily carried out what was defined as “phonological reductions” and “rebuses” (44.6% and 35.6% respectively). This type of modification corresponds with the modifications that we classified in the category of reductions to which the consonant skeletons belonged. These results also corroborate those obtained by C. Combes *et al.* (2012) demonstrating a far greater use of consonant skeletons in expert texters than in novice texters (43.5% vs. 19.5%). The reductions therefore seem to be specific to eSMS. These results show that the expert texters automated a certain type of eSMS modification. The automation of this writing takes root as expertise in eSMS is acquired, thereby explaining why this kind of modification was rarely used by novice texters. However, this difference between novice and expert texters is significant, and imply that expert texters.

Furthermore, the results obtained for the dual-task situation showed the cognitive cost of SMS writing for participants. These results are consistent with a possible automation of the SMS production processes. The results for the “other modifications” observed in the message body and closing showed that the participants produced more modifications of this type in the dual-task situation than in the single-task situation. More precisely, these modifications were of the ‘replacement by another formulation’ and ‘absence’ type and primarily appeared at the end of the dictated message. These results might be explained by the fact that eSMS is cognitively costly for participants. It is possible that, due to their lack of expertise in eSMS, the participants required far more resources in order to write in eSMS and therefore had greater difficulty remembering the end of the dictated message. This may be why the participants replaced the original closing with another more automatic closing or used no closing at all. Writing in eSMS was costly and not automatic for participants, and they were therefore unable to retain the entire message in their memory.

Finally, the results obtained demonstrated an effect of the message section, meaning that the participants had made more modifications in the closing than in the opening as well as in the message body. As these modifications were primarily replacements by another formulation and absences, this implies that the participants had more difficulty keeping the end of the message in their memory than the beginning. These results might be explained by the limited ability of the texters to handle the information they have at their disposal (C. Bereiter and M. Scardamalia, 1987; B. Bourdin and M. Fayol 1994; L.S. Flower and J.R. Hayes 1980).

Based on these findings, it can be concluded that the spelling forms produced in eSMS, and thereby the processes by which they were produced, differ according to the degree of SMS expertise as well as the attention that the participants were able to devote to the task of writing the SMS. For novice texters, this combination of results is compatible with a cognitively costly task of writing in eSMS language, a task which has a tendency to become automatic as the users gain expertise.

New questions also arise from these results regarding how to improve the understanding of the processes involved in producing SMS spelling forms. These questions include the impact of the dual task and the influence of the degree of SMS expertise on the attentional resources required for the SMS spelling production process. So, although these results present the effect of these two factors, it would be worthwhile going further in this research. Indeed, the results that were obtained in this study imply that the secondary task composed of pseudowords had a greater impact on the storage of information in the phonological loop of the WM than the 'beeps' used in the study by C. Combes *et al.* (2012). However, these results imply that the 'verbal' nature of the secondary task used here was responsible for a deterioration of the performance on the primary task. One way to ensure that the secondary task does not interfere with the performance of the primary task is to call upon other WM resources than those of the phonological loop. Indeed, as A. D. Baddeley's (1986) model showed, the phonological loop is not the only aspect to be implicated in the tasks of reading and writing; the visuo-spatial sketchpad also plays its part (A. Dédéyan, P. Largy and I. Negro 2006). It would therefore be interesting to continue this study by proposing a secondary task which implicates the visuo-spatial sketchpad.

Acknowledgements

Special thanks are extending to Martha Randy for translating this paper.

References

- Alamargot, Denis; Eric Lambert; Lucille Chanquoy. 2005. La production écrite et ses relations avec la mémoire. *Approche Neuropsychologique des Acquisitions de l'Enfant* 17: 41–46.
- Anis, Jacques. 2001. *Parlez-vous texto? Guide des nouveaux langages du réseau*. Paris: Le Cherche Midi.
- Anis, Jacques. 2002. Communication électronique scripturale et formes langagières: Chats et SMS. Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques. Université de Poitiers, 31 mai et 1er juin 2002: 57–70.
- Baddeley, Alan D. 1986. *Working Memory*. New York: Oxford University Press.
- Baddeley, Alan D. 1992. Working memory. *Science* 255(5044): 556–559.
- Bereiter, Carl; Marlene Scardamalia. 1987. *The Psychology of Written Composition*. Hillsdale, NJ: Erlbaum.
- Berninger, Virginia W.; H. Lee Swanson. 1994. Modification of the Hayes and Flower Model to Explain Beginning and Developing Writing. In *Advances in Cognition and Educational Practice. Children's Writing: Toward a Process Theory of Development of Skilled Writing*, E. Butterfield (ed.), Vol. 2, 57–82, Greenwich, CT: JAI Press.
- Bourdin, Béatrice. 1999. Mémoire de travail et production langagière: comparaison de l'oral et de l'écrit chez les adultes et les enfants. *L'année psychologique* 99(1): 123–148. doi: 10.3406/psy.1999.28551
- Bourdin, Béatrice; Michel Fayol. 1994. Is a written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology* 29(5): 591–620.
- Brissaud, Catherine; Jean-Pierre Chevrot. 2001. Acquisition de la morphographie entre 10 et 15 ans: le cas du pluriel des formes verbales en /E/. *VERBUM* XXII(4): 425–439.
- Combes, Céline; Olga Volckaert-Legrier; Pierre Largy. 2012. Différences novices — experts dans la production écrite de SMS: étude de l'effet d'une double tâche. *Approche Neuropsychologique des Acquisitions de l'Enfant* 24(119): 302–312.
- Crystal, David. 2001. *Language and the Internet*. Cambridge: CUP.
- Dédéyan, Alexandra; Pierre Largy; Isabelle Negro. 2006. Mémoire de travail et détection d'erreurs d'accord verbal: étude chez le novice et l'expert. *Langages* 4(164): 57–70. doi: 10.3917/lang.164.0057
- Doutriaux, F.; R. Lepez. 1980. *Test de Niveau d'Orthographe (TNO)*. Paris: ECPA.
- Elliot, Collin D.; Pauline Smith; Kay McCulloch. 1996. *British Ability Scales: Second Edition (BAS II)*. Windsor: NFER-Nelson.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier. 2006a. Le langage SMS: révélateur d'1 compétence. In *"Le français m'a tuer": actes du colloque L'orthographe à l'épreuve du supérieur: Institut libre Marie Haps, Bruxelles, 27 mai 2005*, J. J. Didier, O. Hambursin, P. Moreau, M. Seron (eds.), 33–41. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier. 2006b. *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciel de consultation* [CD-Rom]. Louvain-la-Neuve: Presses universitaires de Louvain.
- Fayol, Michel; Pierre Largy. 1992. Une approche cognitive fonctionnelle de l'orthographe grammaticale. *Langue Française* 95(1): 80–98.

- Fayol, Michel; Pierre Largy; Patrick Lemaire. 1994. When cognitive overload enhances subject-verb agreement errors. A study in French written language. *Quarterly Journal of Experimental Psychology* 47A: 437–464.
- Fayol, Michel; Ania Miret. 2005. Ecrire, orthographier et rédiger des textes. *Psychologie française* 50(3): 391–402. doi: 10.1016/j.psf.2005.05.008
- Flower, Linda S.; John R. Hayes. 1980. The Dynamic of Composing: Making Plans and Juggling Constraints. In *Cognitive Processes in Writing*, L.W. Gregg, E.R. Steinberg (eds.), 31–55, Hillsdale, NJ: Erlbaum.
- Hayes, John R. 1996. A New Framework for Understanding Cognition and Affect in Writing. In *Cognitive Processes in Writing*, C. M. Levy, S. Ransdell (eds.), 3–30, Hillsdale, NJ: Erlbaum.
- Hayes, John R.; Linda S. Flower. 1980. *Identifying the Organization of Writing Processes*. *Cognitive Processes in Writing*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kellogg, Ronald T. 1987. Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition* 15: 256–266.
- Kellogg, Ronald T. 1994. *The Psychology of Writing*. New York: Oxford University Press.
- Kellogg, Ronald T. 1996. A Model of Working Memory in Writing. In *The Science of Writing*, C. M. Levy, S. Ransdell (eds.), 57–72, Mahwah, NJ: Erlbaum.
- Largy, Pierre; Marie-Paule Cousin; Alexandra Dédéyan. 2005. Produire et réviser la morphologie flexionnelle du nombre: de l'accès à une expertise. *Psychologie française* 50: 339–350.
- Liénard, Fabien. 2008. Analyse linguistique et sociopragmatique de l'écriture électronique de l'écriture électronique. Le cas du SMS tchaté. In *La langue du cyberspace: de la diversité aux norms*, J. Gerbault (ed.), 265–278, Paris: l'Harmattan.
- McCutchen, Deborah. 1996. A capacity theory of writing: Working memory in composition. *Educational Psychology Review* 8(3): 299–325.
- Mourlhon-Dallies, Florence. 2010. Modifications et inventions graphiques dans les écritures électroniques. *Le Français d'aujourd'hui* 3(170): 101–112. doi: 10.3917/lfa.170.0101
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, T. Arnavielle (ed.), 33–52, Université Paul-Valéry: Montpellier 3.
- Plester, Beverly; Clare Wood; Victoria Bell. 2008. Txt msg n school literacy: Does texting and knowledge of text abbreviations adversely affect children's literacy attainment? *Literacy* 42(3): 137–144. doi: 10.1111/j.1741–4369.2008.00489.x
- Power, M. J. 1985. Sentence production and working memory. *Quarterly Journal of Experimental Psychology* 37A: 367–385. doi: 10.1080/14640748508400940
- Schneider, Walter E.; Richard M. Shiffrin. 1977. Controlled and automatic human information processing. I. Detection, search, and attention. *Psychological Review* 84(1): 1–66.
- Traverso, Véronique. 1999. *L'analyse des conversations*. Paris: Nathan.
- Wood, Clare; Emma Jackson; LaVaughn Hart; Beverly Plester; Larry Wilde. 2011. The effect of text messaging on 9- and 10-year-old children's reading, spelling and phonological processing skills. *Journal of Computer Assisted Learning* 27: 28–36. doi: 10.1111/j.1365–2729.2010.00398.x
- Wood, Clare; Emma Jackson; LaVaughn Hart; Beverly Plester; Larry Wilde. 2011. The effect of text messaging on 9- and 10-year-old children's reading, spelling and phonological processing skills. *Journal of Computer Assisted Learning* 27: 28–36. doi: 10.1111/j.1365–2729.2010.00398.x

Summary

The objective of this study was to attempt to distinguish the various processes of producing SMS spelling forms. The production of these different spelling forms was compared by means of an experimental paradigm: the dual task. This paradigm aimed at identifying the attentional resources necessary for the process of producing SMS spelling. Another way in which to address the degree of automation of these production processes was to compare SMS productions in terms of the level of SMS writing expertise. The results of this study demonstrated that the spelling forms produced in SMS language (eSMS), and therefore their production process, differ according to the degree of SMS writing expertise and the attention that the participants are able to devote to the SMS writing task. The results confirm that SMS writing represents a cognitive cost for novice texters and tends to become automatic as the users acquire expertise.

Development of SMS language from 2000 to 2010

A comparison of two corpora

Úrsula Kirsten Torrado
Universidade de Vigo

Introduction

The new technology era doubtless has an impact on the way people communicate with each other: via telephones, mobile phones, e-mails, chats and instant messages among others (see N. Baron 1998; D. Crystal 2001; C. Thurlow and S. McKay 2003; C. Frehner 2008), especially teenagers, who use this new technology as a way of expressing themselves in a modern style. However, SMS language is used in every age range and social status. Undoubtedly, text messaging or short-message service (hereafter, SMS) communication has become progressively widespread as a social phenomenon owing to its effectiveness and efficiency and because it is “a quick, cheap and easy to use” way of communication that more and more people use every day (R. Grinter and M. Eldridge 2001: 219).

The immediacy of SMS has helped to spread this “new” communication system (cf. T. Shortis 2007; D. Crystal 2009; M. Markus 2010); in fact, many teenagers prefer to let their fingers do the talking than to engage in a face-to-face conversation (S. Kunda 2005).

It may be thought that the 12-key keypad, the pocket-size screen and the small amount of 160 characters, including letters, numbers or symbols from the Latin alphabet, has caused a number of adaptations of spelling when texting. However, C. Thurlow’s research (2003) proves that the length of text messages is shorter than the one allowed when sending SMSes (160 characters including spaces). Usually the space available is not fully used by the texters. Thus, the spelling adaptations would be caused by “the needs for speed, ease of typing and, perhaps, other symbolic concerns” (C. Thurlow 2003: 4). In addition, the research carried out proves that some words are lengthened to emphasize accent, slang and attitude. Thus, words are not always shortened as could be expected.

This paper is an attempt to contribute to this recent area of study. The purpose is to determine the development of SMS from 2000 to 2010 and to investigate the relationship that English native speakers have established between English pronunciation and SMS language. The rise of this type of SMS language is called “Textspeak” by D. Crystal (2003, 2008b). In order to achieve this purpose a comparison of two corpora dating from 2000 and 2010 respectively were used. This research focuses on determining the features employed in those SMSes (features adapted from C. Thurlow, 2003).

The initial working hypothesis on which this study is based is that phonologically-based devices are expected to be more common than other devices such as contractions or initialisms, as it is easier to understand shortened forms when there is a sound-to-spelling relationship and both words are homophonous or near homophones.

1. The two corpora

For the purpose of the diachronic analysis of this paper two different corpora were employed: (a) a free online British SMS corpus compiled in 2000, and (b) my own data compiled by means of questionnaires in 2010.

1.1 Description of the online corpus

A free online British SMS corpus which was previously available at netting-it.com containing 202 text messages compiled in 2000 was used for this research. It was compiled as a sample corpus to accompany T. Shortis’s textbook *The language of ICT- Information and Computer Technology* (2001). Due to the lack of information available online, Dr Tim Shortis was contacted in order to obtain more information about the corpus employed for the research. The corpus was taken from a group of about 50 UK college pupils and their family and friends in 2000. The pupils were all 17–18 year-olds from a sixth form college in Bristol (St Brendan’s Sixth Form College). Sixth form colleges are educational institutions in England and Wales where pupils study to achieve advanced school-level qualifications such as A-levels to continue with university studies. They are also known as junior colleges. They are relatively academic colleges for pre-university pupils who choose to stay on after compulsory schooling. According to T. Shortis (personal communication), the original corpus was over 600 messages but it was edited down as some subjects could not be contacted for permission for the messages to go on the Internet. In order for the participants to remain anonymous, names were changed by Jon Stevenson.

Microsoft Word was used to work out some statistics and patterns in the corpus with the 'word count' function. The length of each individual message was calculated to obtain the overall results.

The corpus consists of 3,957 words in total although the calculation is somewhat complicated as

- a. in some text messages there are two or more words combined which stand for two or more words but are counted as one (e.g. '2U' meaning 'to you');
- b. there are also examples of the opposite: several words which stand for just one word but count as several (e.g. 'n e thin' meaning 'anything');
- c. there are cases of several words together without any space in between which are therefore counted as just one word although they stand for several ones (e.g. 'tellmiss' instead of 'tell miss'). This may be a mistake made by the writer but the results vary with this type of example.
- d. Sometimes spaces are left where they should not be, as for example in between exclamation marks or hyphens. Thus, they will be taken as two words instead of one (e.g. 'mate! !' instead of 'mate!!'). In addition to these four aspects, there are other situations that can be mentioned:
- e. Some apostrophes are missing from the contractions (e.g. 'ill' instead of 'I'll', meaning I will). These are devious examples which, as when contractions are used with the apostrophe, also count as one word even if these contractions stand for at least two words.
- f. There are also repetitions of certain words which are probably mistakes made by the writers, maybe due to forgetfulness when writing SMSes or because they did not review the message before sending it. There are also other types of mistakes within some words where the spelling is changed or another letter is added, maybe due to the previous reasons or maybe due to typing mistakes. Although the number of words may or may not be increased, the number of characters per message will definitely change (e.g. 'please dont please dont ignore my calls' instead of 'Please don't ignore my calls' or 'whill' instead of 'will').

On this basis, the entire number of characters including spaces is 18,328 and excluding spaces it is 14,565. There are two text messages which consist of only one word:

- a. Message number 87: *ALRITE*
- b. Message 128: *G.W.R*

The longest message is message number 161, which consists of 41 words:

- c. *'Hey! do u fancy meetin me at 4 at cha — hav a lil beverage on me. if not txt or ring me and we can meet up l8r. quite tired got in at 3 v.pist ;) love Pete x x x'*

The average message length in the corpus is 19.58911 words. The message with the smallest amount of characters excluding spaces consists of three characters (message 89: 'U 2.'), and the message with the highest amount of characters excluding spaces consists of 139 characters (message 186: 'Si.como no?!listened2the plaid album-quite gd&the new air1 which is hilarious-also bought"braindance"a comp. ofstuff on aphex's;abel,u hav2hear it!c u sn xxxx'). The average number of characters per message is 72.10396.

The corpus is not tagged and, therefore, there is no information about the writers of these text messages, such as their gender. We only know the age of the participants (17–18 year-olds) which could be guessed through the type of messages and the language employed which suggests that the writers are (young) adults or at least that they are not teenagers (people under 17 years). Further research could aim to prove whether or not there is any difference in the SMS language employed by teenagers and adults.

1.2 Compilation of my own data

For the compilation of my own data, three different questionnaires were designed: pupils' questionnaire part A (hereafter, SQ1), pupils' questionnaire part B (hereafter, SQ2) and teachers' questionnaire (hereafter, TQ). In the first questionnaire pupils were asked for some personal information and some information about their usage of mobile phones. In the second questionnaire pupils were asked to transcribe different text messages. In the third questionnaire teachers were asked to give their impressions on the impact that mobile phones have on their pupils' message writing and on other types of writing.

For the purpose of this analysis, only SQ2 is relevant. This questionnaire was filled in by 50 subjects who participated in the research. There were approximately the same number of males (22) and females (28) as it is necessary to check whether the data changes depending on the gender of the subjects. Subjects were mainly 13 and 14 years old, although there were four 12 year-olds. Participants were studying year 8 and 9 of Key Stage 3 (hereafter KS3).¹

Questionnaires were carried out in their schools. I handed out SQ2 to subjects in May 2010 myself. Pupils needed more than one registration period and some teachers from the school were kind enough to loan some of their lessons so that

1. KS3 refers to the first three years of Secondary Education (years 7, 8 and 9) in England and Wales, when students are between 11 and 14 years old.

pupils were able to fill in the questionnaires during normal lessons. Unexpectedly, pupils required an average of one hour and thirty minutes to fill in the questionnaires. The same pupils completing SQ1 did SQ2. A pilot study was carried out in order to check the reliability of this questionnaire and to ascertain whether any modifications were required. The results of the pilot study suggested the need to add a question about the subjects' usage of predictive text in SQ2.

Moreover, SQ2 allowed me to create my own SMS corpus, which will serve as the basis for analysing and processing the data in the diachronic analysis of SMS language for the last 10 years.

1.2.1 *Authorisation*

As the respondents of the questionnaires were pupils under the age of 18, in order to follow the procedures for child protection, it was decided that their guardians would be required to sign an authorization to show their agreement with the participation in the research, *i.e.*, their guardians were subject to formal written consent. Although the consent had not been imposed either by the institution they belong to or by the investigator's institution, as the research concerns subjects under the age of 18, it was a way to protect them and to assure the guardians that all the data provided was going to be used in an ethical, responsible and sensitive way. Moreover, the authorization guarantees the confidentiality of the respondents' identities. However, very few pupils managed to return the authorisation signed, probably due to forgetfulness. Thus, an alternative authorisation was created, where pupils were required to sign the authorisation where they stated that the data provided could be used by the researcher for statistical purposes.

1.2.2 *Pupils' questionnaire part A (SQ1)*

At the beginning of the questionnaire there is a brief introduction explaining the purpose of the questionnaire and introducing myself. It is also emphasized that the subjects' identities will not be disclosed and will be kept confidential. Although at a first sight the questionnaire may seem quite long, it is actually short and takes approximately 10 minutes to complete.

This questionnaire contains a mixture of closed-response items and open-response items, as the classification that James Dean Brown (Heigham and Broker 2009:200–216) suggests for the types of questions that can be found in questionnaires; although the closed-response items are more frequent, there are several open-response items. The questions that use a closed-response structure are the following: question 2, question 14 to 23, 25, 26, 28, 29 and question 31 to 33. The questions using an open-response structure are question 1, question 3 to 13, question 24 to 27 and question 30.

In this first questionnaire, the first 16 questions are biodata items. Respondents were required to give some personal information about themselves. All those questions except numbers 2, 14, 15 and 16 are fill-in items that are a type of open-response items. The subjects provided personal information that was considered relevant for the research, such as first languages at home, place of birth, nationality or school, among other information. Questions 2, 14, 15 and 16 are closed-response items. In question 14 respondents were asked about disabilities. If they have any disability they need to go to question 15 and specify which one. The disabilities mentioned in the questionnaire are the ones that might have consequences for the subjects' spelling and, therefore, the results of those pupils would not be accurate. On the other hand, in question 16, the subjects are asked about their ethnic background. As there are sometimes linguistic differences depending on the ethnic background, it will be checked if there is any variation in the way text messages are written, although my hypothesis is that the ethnic background will not be relevant.

Questions 17 to 32 all deal with mobile phones. Those questions ask about the possession of mobile phones, their use and other impressions related to their use. Questions 24 and 27 are specific open questions of the short-answer items type, which is a type of open-response item. Question 30 is a type of sentence completion item.

In question number 17, participants are asked whether or not they have their own mobile phone, since if they do not have their own mobile phone, the results may vary, as they would send fewer text messages. Questions 18, 19 and 20 ask about their usage. Questions 21 to 27 will provide information about the impression subjects have of the influence text messages have on their writing. I do think new technologies have an influence on the way teenagers write, and those questions will help me check whether or not they have the same impression. Questions 28 and 29 are designed to obtain information about the influence of the addressee on the way text messages are written. Question 30 is important to check whether native speakers find it easy to understand the new SMS language. Question 31 will be helpful to determine whether writing that way is a trendy way of expressing themselves. Question 32 is very important as the present research will deal with the sound-to-spelling relationship in SMS language. I do think there is an undoubted relationship but I would like to check whether or not native speakers have the same impression as I do. In question 33 participants are asked to indicate which features of the ones in the list they use when sending text messages. The list is adapted from C. Thurlow, 2003.

1.2.3 *Pupils' questionnaire part B (SQ2)*

The beginning of the questionnaire contained a brief introduction explaining the purpose of the questionnaire and a section where I introduce myself. It was also emphasized that the subjects' identity will not be disclosed and will be kept confidential.

In the pupils' part B questionnaire I will focus on the aspects that are relevant for the relationship between spelling and pronunciation, which are: clippings, letter/number homophones, misspellings and typos, non-conventional spellings and accent stylisation. Obviously the use of function words and punctuation will be noted. Once the pupils had completed the questionnaire, the data was analysed and it was possible to check if there is any relationship between the abbreviations used by the subjects and their pronunciation.

In this questionnaire pupils were asked to turn 30 sentences into text messages. Students were expected to transcribe the sentences given, to check if they were using any of the devices or spelling adaptations described in the following section. Nonetheless, they actually modified the sentences by changing some of the sentences or words given by others they would use instead. Therefore, sentences such as: *'How are you today?'* were replaced by *'WUU2'* ('What are you up to?'). Thus, sentences were somehow changed into similar sentences that sounded more natural to them. Moreover, some words or expressions were added (see Section 2.4: Other significant features). Participants used colloquial expressions and words that were not in the original sentences. Although the meaning of the sentence remains the same, the way of expressing the message varied depending on the participant.

Participants were also asked about the smileys they use in their personal mobile communication and about the use of predictive text when texting other people.

2. Comparison of the data

2.1 Classification of spelling adaptations

Unlike English standard spelling, SMS spelling is heterogeneous and uses a wide range of different devices. Different classifications have recently been proposed. C. Thurlow (2003:4) classified the following spelling adaptations: shortenings, contractions, g-clippings and other types of clippings, acronyms, initialisms, letter/number homophones, 'misspellings' and typos, non-conventional spellings and accent stylisation. However, the list can be increased by adding other features such as smileys, lack of function words, lack of punctuation or over-punctuation. Several of those features imply a relationship between English pronunciation and

spelling in SMS language. Nevertheless, it must be born in mind that, as stated by T. Shortis (2007) and D. Crystal (2008a; 2009), texting is not the cause of bad spelling and it can improve the user's literacy. Hence, statements claiming that the new non-standard spelling implies language decay might be false, although it seems to be a widespread idea among the general public to consider that SMS language has a negative effect on language. Actually, in the TQ, out of 24 subjects, 19 (79.2%) support the idea that SMS language has a bad influence on their pupils' writing and only 5 (20.8%) think that it has a neutral influence. However, some researchers (T. Shortis 2007; D. Crystal 2008a; 2009), suggest that texting is not the cause of bad spelling and it can improve the user's literacy regardless of the popular belief of its bad influence; people using SMS language need to know the actual spelling of words before making any spelling change that depicts this type of language.

D. Crystal (2009) has offered a classification of spelling adaptations into pictograms and logograms (rebus abbreviation), initialisms, omitted letters (contractions and clippings), nonstandard spellings, shortenings and genuine novelties.

P. López-Rúa (2007) classifies the devices employed in text messages into abbreviations, clippings, initializations (acronyms, alphabetisms and abbreviations composed of initials), phonetic respellings, letter and number homophones, and symbols and onomatopoeic expressions.

In this research the following classification will be followed:

1. Reductions: All types of shortened forms not based on phonology.
2. Phoneticised respellings: All types of shortened forms based on the relationship between English pronunciation and spelling.
3. Word deletion and punctuation: Absence of grammatical words and usage or non-usage of punctuation in SMS language.
4. Smileys and other symbols, also known as 'emoticons' (D. Crystal 2004: 38). "A sequential combination of keyboard characters designed to convey the emotion associated with a particular facial expression," either positive or negative. "Emoticons are typed as a string on a single line, and usually located at the end of a sentence; most need to be read sideways." (D. Crystal 2004: 38–39).

As for the devices employed throughout this paper for the classification of spelling adaptations used when texting, C. Thurlow (2003)'s devices were adapted with some variations as follows:

1. Reductions:
 - 1.1. Shortenings (sometimes called 'back clipping' or 'apocopation'): dropping end letter(s). E.g. *hav* ('have').
 - 1.2. Contractions (also called 'middle clipping' or 'syncope'): dropping middle letter(s). E.g. *txt* ('text').

- 1.3. Other types of clippings (also known as ‘fore-clipping’ or ‘aphaeresis’): missing first letter(s). E.g. *morrow* (‘tomorrow’).
- 1.4. G-clippings: loss of final -g. E.g. *darlin* (‘darling’).
- 1.5. Acronyms: words consisting of more than just initials pronounced as one word. Words will be classified into acronyms or initialisms depending on their orthographic shape and the way in which they are read. Therefore, if a word consists of just initials and they are pronounced one by one it will be considered as an initialism, but if it consists of more than just initials and it is pronounced as one word it will be considered as an acronym. E.g. *NATO* (‘North Atlantic Treaty Organization’). It must be borne in mind that these classifications may vary depending on the source used. Thus, some authors distinguish between acronyms, abbreviations, initialisms and alphabetisms (G. Cannon 1989, P. López- Rúa 2002 and 2004).²
- 1.6. Initialisms: words consisting of just initials pronounced one by one. E.g. *NHS* (‘National Health Service’) or *v* (‘very’).
- 1.7. Misspellings and typos: misprints or spelling mistakes not performed intentionally as a device or as a spelling adaptation. E.g. *this* (‘this’).
- 1.8. Abbreviations (other types of shortened forms not included in the previous devices): affectionate forms of names and other words shortened where the word is shortened but there is at least one letter not included in the full word (eg. *no* ‘number’).

2. Phoneticised respellings:³

2. P. López-Rúa (2004:36) classifies initialisms into acronyms and alphabetisms, claiming that they “are ‘degree of shortening’, ‘source form’, and ‘pronunciation’: in this way, acronyms tend to be straightforwardly described as words formed from the initial letters or larger parts of the words in a phrase, and the parameter ‘pronunciation’ is often used to distinguish them from alphabetisms.”

As stated by G. Cannon (1989: 107), acronyms are a subclass of abbreviations or vice versa. An acronym is described as:

‘[A] word formed from the initial letter or letters of each of the successive parts or major parts of a compound term’ (G. Cannon, 1989: 107). ‘An acronym preserves only the initial part(s) of a multiword source [...] traditional acronyms preserve at most only two initial letters/sounds [...] [A]n acronym must come from a source with at least three constituents, where a combining form can be a constituent (ASP “Anglo-Saxon Protestant”). Not more than two initial letters/sounds of some or all of the constituents can be retained, though an exception of three or even four is permitted if the majority of the reduction typifies acronymy.’ (G. Cannon, 1989: 108)

3. Initially phoneticised respellings were divided into three different devices; after the analysis of SQ2 it was necessary to add a fourth device named ‘stylish talk’. This device has not been used in the analysis of the SMS corpus as no examples were found and it was added *a posteriori*.

- 2.1. Letter/number homophones: a word or part of a word is replaced by a number or letter with the same pronunciation. E.g. *u* ('you') or *2* ('to, too, two').
- 2.2. Non-conventional spellings: abbreviation of a word respelling it phonetically, *i.e.*, they are phonetic respellings. E.g. *luv* ('love').
- 2.3. Accent stylization: 'phonological approximation such as the "regiolectal"' (J. Androutsopoulos 2000: 521, quoted in C. Thurlow 2003: 4) and onomatopoeic and exclamatory expressions. E.g. *da* ('the').
- 2.4. Stylish talk: Similar to accent stylization but words are lengthened in size. Thus, the resulting word is longer than the standard one. E.g. *Hugsz* ('hugs').

Apart from the devices mentioned above, the following devices have also been added as they are also employed quite frequently in text messages:

3. Word deletion and punctuation:
 - 3.1. Lack of function words: loss of grammatical words. E.g. *you want to* ('do you want to').
 - 3.2. Lack of punctuation: loss of punctuation. E.g. *he hasnt* ('he hasn't').
 - 3.3. Over-punctuation: excessive use of punctuation. E.g. *why???* ('why?').
 - 3.4. The use of capital letters: use of capitals, lower case or both. E.g. *yeP* ('yep').
4. Smileys and other symbols. E.g. *xxx* ('kisses').

For the sub-categorisation, the devices proposed by C. Thurlow (2003) have been preferred over D. Crystal's and P. López-Rúa's, as they provide a much more detailed classification and therefore more information can be provided on the different types of words employed throughout the corpus used for the research.

2.2 Coding of the databases

In order to process and analyse the data obtained in both corpora, two databases were created, one per corpus. The SMSes from the OC and MOC (202 and 1500 text messages respectively) were analysed manually, one by one. Every word illustrating any spelling adaptation was manually classified into the different categorization of devices explained in detail in Section 2.1. Each of the words classified was considered a token, that is to say, a case in point of the device. Some of the instances emerged repeatedly, which increased the number of tokens, but remained as one type. Thus, the letter homophone '*u*' ('*you*') is one type that came out 642 times in MOC and, therefore, there are 642 tokens. We can see an example of the g-clippings from MOC. There are a total of 170 tokens and 20 types. We can also see that some of the words are highlighted in bold, because those words have also

been classified as another device. For instance, the word ‘*sk8in*’ (‘*skating*’) has been classified as a g-clipping and as a letter/number homophone. Thus, that word will count as two tokens. The following table is an example:

Table 1. G-clippings in MOC

G- clippings	Abs. Freq	Rel. Freq
breathin = breathing	1	0.59
buyin = buying	2	1.18
comin = coming	3	1.76
dippin = dipping	1	0.59
doin = doing	24	14.12
goin = going	69	40.59
gwoin = going	1	0.59
letin = letting	2	1.18
lettin = letting	16	9.41
sayin = saying	5	2.94
sk8in = skating	4	2.35
sk8Tin = skating	1	0.59
studin = studying	1	0.59
studyin = studying	11	6.47
tellin = telling	2	1.18
w8in = waiting	2	1.18
waitin = waiting	13	7.65
watin = waiting	1	0.59
wishin = wishing	1	0.59
TOTAL	170	100

2.3 Devices used in both corpora

In view of the typology of spelling adaptations adopted for this study and explained in Section 2.1., the following table shows the results of the adaptations used in both corpora analysed for this paper: (a) the Online Corpus which used to be at netting-it.com (hereafter: OC), and (b) my own corpus processed after the analysis of SQ2 when transcribing into SMS language 30 text messages (hereafter, MOC).

Table 2. Devices used in SMS. Comparing OC and MOC (2000 to 2010)

SMS LANGUAGE	Tokens		Types	
	OC (2000)	MOC (2010)	OC (2000)	MOC (2010)
Reductions	379 (30.5%)	2,085 (33.53%)	216 (48.97%)	433 (48.5%)
Phoneticised respellings	737 (59.3%)	3,928 (63.16%)	201 (45.57%)	436 (48.8%)
Smileys and other symbols	126 (10.1%)	206 (3.31%)	24 (5.44%)	24 (2.7%)
Total	1,242 (100%)	6,219 (100%)	441 (100%)	893 (100%)

According to the data processed, there are a total of 1,242 tokens in the OC and 6,219 tokens in MOC. However, it should be taken into consideration that some of the examples have been classified as two different devices as they may contain two different features; therefore, they have been counted as two tokens. The important difference in the number of tokens between the OC and MOC is due to the fact that the OC consists of 202 SMS and MOC consists of 30 SMS per subject, which amount to a total of 1,500 text messages. This is the main reason for the difference in the total amount of tokens.

Although the usage or absence of punctuation and capital letters has its relevance when dealing with SMS language, they will not be considered in this analysis as those aspects are not purely linguistic. As regards smileys, in the OC the number is not markedly high, and in MOC they were used more often than expected, as there were no indications for the pupils to add them. This means that pupils were adding them whenever they felt they were appropriate, even when the text messages were not specifying any sort of emotion. Actually, in MOC emoticons were not expected at all. According to the previous table, the most common adaptations when texting are phoneticised respellings (about 60% of the examples found in the OC and 63.16% of the examples found in MOC), which is practically twice the number of reductions (about 30% of the examples in the OC and 33.53% of the examples found in MOC). This is quite a remarkable fact, as there are eight different devices in reductions whereas phoneticised respellings are only sub-classified into four devices, which are half the devices. This is statistically significant as in proportion, we would expect reductions to be more common.

Nevertheless, from a more in-depth point of view, if repeated examples are not taken into account, it is noteworthy that in the case of the OC, the number of reductions is more common than the cases of phoneticised respellings. There are 216 examples of reductions (48.97%) and 201 examples of phoneticised respellings (45.57%). Regarding MOC, there is almost the same amount of reductions (48.5%) and phoneticised respellings (48.8%).

Bearing in mind the total number of examples, including repetitions, there is a greater number of examples of phoneticised respellings. This is quite a

thought-provoking phenomenon, as it indicates that phoneticised respellings are a very recurrent device and it also proves the initial hypothesis that phonologically-based adaptations are more frequent than the other ones; not surprisingly, the most frequent adaptations are the ones based on pronunciation. The main reason might be for communication purposes: if two words with different spellings are homophonous, and therefore have the same pronunciation, comprehension will be less problematic and complicated than if the text message is ambiguous. For instance, initialisms and acronyms are quite difficult to understand if the reader does not know what the message is about, and that may hinder understanding.

In addition, it might be claimed that native speakers, following the popular belief that spelling is odd and hilarious, have somehow developed the ability to find connections between English spelling and pronunciation. Hence, native speakers are able to find an alternative spelling for common words with the same pronunciation. This may show that there is an intrinsic relationship between English pronunciation and spelling regardless of the independence that could be expected at first sight. It can be assumed that the adaptations based on the sound-to-spelling relationship, *i.e.*, phonological approximation or phoneticised respellings, are more frequent than other types of adaptations. The reasons for that might be the lack of ambiguity in those types of examples, as they are homophonous words, and also due to speed and, perhaps, brevity. Although in our analysis of the OC brevity would be a clear reason, it is not as clear in a more recent corpus (MOC) dating from 2010. We should acknowledge that a new device seems to be currently taking place in SMS language. Hence, some words are respelled in such a way that the resulting word is longer than the standard one, which would be a trend against maximum brevity. This is an outstanding fact, as SMS language is usually understood as a way of rewriting words in a shortened way. In the majority of instances those words could be classified as accent stylization, as they try to imitate the way people talk. However, some words have not been classified under this device as the resultant word is longer than the original form.⁴ Thus, those words have been classified into a new device, which has been named 'stylish talk'. Some of the examples found are *lotsz* ('lots'), *friendsz* ('friends'), *tuu* ('to'), *byee* ('bye'), or *today* ('today').

On balance, it might be said that SMS language seems to have evolved during recent years. It seems to be turning into a fashionable and stylish way of writing, where shortened versions of the words are not always the aim of the respelling. Further research should be carried out to check whether this happens in other areas outside London and in all age ranges.

The following table displays the devices used for reductions.

4. Only onomatopoeic and exclamatory expressions have been classified as accent stylization regardless of the length of the resultant word.

Table 3. Devices used for reductions. Comparing OC and MOC (2000 to 2010)

REDUCTIONS	Tokens		Types	
	OC (2000)	MOC (2010)	OC (2000)	MOC (2010)
Shortenings / Apocopation	124 (32.7%)	341 (16.4%)	64 (29.6%)	90 (20.8%)
Contractions / Syncope	89 (23.5%)	1,224 (58.7%)	52 (24.1%)	236 (54.5%)
G-clippings	76 (20.1%)	170 (8.1%)	36 (16.7%)	20 (4.6%)
Other types of clippings / Aphaeresis	5 (1.3%)	67 (3.2%)	5 (2.3%)	14 (3.2%)
Abbreviations	24 (6.3%)	31 (1.5%)	11 (5.1%)	10 (2.3%)
Acronyms	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Initialisms	31 (8.2%)	225 (10.8%)	21 (9.7%)	45 (10.4%)
“Misspellings” and typos	30 (7.9%)	27 (1.3%)	27 (12.5%)	18 (4.2%)
Total	379 (100%)	2,085 (100%)	216 (100%)	433 (100%)

Although the most frequent adaptations are those based on pronunciation, contractions and shortenings are also frequently used. As the previous table shows, in the OC the most common reduction device is shortenings and in MOC the most frequent reduction device is contractions. The order of frequency of the devices used in reductions in the OC is as follows: shortenings, contractions, g-clippings, initialisms, misspellings and typos, abbreviations, other types of clippings and acronyms. In the OC the omission of letters in words is the most common practice. Hence, the dropping of final letters (*i.e.*, shortenings), the omission of middle letters (*i.e.*, contractions) and the dropping of final -g (*i.e.*, g-clippings) are the most common reduction devices, respectively. The only uncommon clipping is the dropping of the first letter. The reason may be that it seems easier to recognise a word if the ending is missing than if the beginning of the word has been dropped.

The following most frequently used reduction devices are initialisms, ‘misspellings’ and typos, abbreviations and other types of clippings, all of them with less than 10% frequency. There are no acronyms in the corpus. The most likely reason for the low frequency of initialisms and acronyms is that both are quite difficult to understand if the reader does not know what the message is about, and that may hinder understanding. The same may apply to shortenings, contractions, other types of clippings and abbreviations, which are usually easy to understand but may sometimes become difficult. This factor may account for the low frequency of contractions and other types of clippings.

Many instances of shortenings (37.10%), abbreviations (87.50%) and initialisms (48.39%) were proper names, which presupposes the sender’s and addressee’s shared knowledge of those people and, consequently, justifies a somewhat greater usage of shortenings. G-clippings are also quite frequent and the reason may be

that they are also phonologically based as the final *-g* is never pronounced in the gerund *-ing* even though it influences the way in which the *n* is pronounced and, as mentioned before, the devices based on pronunciation are more common. As a matter of fact, if proper names had not been included in shortenings, then the number of shortenings would be cut down to 78 examples; if other capitalised words such as the days of the week were discounted, there would be only 66 examples of shortenings. In both cases, g-clippings would be more common than shortenings.

The order of frequency of the devices of reductions in MOC is: contractions, shortenings, initialisms, g-clippings, other types of clippings, abbreviations, misspellings and typos, acronyms. In MOC this large number of contraction cases may be due to the fact that in this analysis grammatical contractions were included as contraction examples, whereas in the OC they were not, as those words are grammatically correct and exist in standard language. However, in this analysis, they have been included only when pupils were given the subjects and verbs or auxiliaries and the negative particle as two words and pupils transcribed them as one word. If they were already given a contraction, it was not included as an example of SMS language. There is less than 10% frequency in the following devices: g-clippings, other types of clippings, abbreviations, acronyms and misspellings. This coincides with the OC results. The only exceptions are g-clippings, — with a frequency of 20.1% in the OC and 8.1% in MOC —, and initialisms, — with a frequency of 8.2% in the OC and 10.8% in MOC, which is over 10%. The reason for the low frequency of g-clippings may have its root in the low frequency of words ending in *-ing* in the messages that had to be transcribed by the participants (only 8 words containing *-ing*). The slight difference in the frequency of initialisms might be due to the increase in the use of SMS language and to the fact that it has become more common, although the number is still considerably low.

A detailed table with the phoneticised respellings follows:

Table 4. Devices used for phoneticised respellings. Comparing OC and MOC (2000 to 2010)

PHONETICISED RESPELLINGS	Tokens		Types	
	OC (2000)	MOC (2010)	OC (2000)	MOC (2010)
Letter / number homophones	467 (63%)	2,237 (57%)	83 (41.3%)	98 (22.5%)
Non-conventional spellings	144 (20%)	695 (17.7)	54 (26.9%)	117 (26.8%)
Accent stylisation	126 (17%)	908 (23.1%)	64 (31.8%)	163 (37.4%)
Stylish talk	0 (0%)	88 (2.2%)	0 (0%)	58 (13.3%)
Total	737 (100%)	3928 (100%)	201 (100%)	436 (100%)

As can be observed in Table 4, the most common adaptation when texting is letter/number homophones, which almost doubles the number of examples of accent stylization and non-conventional spellings, followed a long way behind by stylish talk. The great number of letter and number homophones probably depends on the lack of ambiguity of the examples, as they are homophonous words. Stylish talk is not very frequent, and the main reason might be that it is quite a new device; in fact there are no examples in the corpora dating from 2000. Therefore, further studies should be carried out to check whether its usage is increased or the same all over the UK.

Regarding stylish talk, it seems to consist of the use of *-ii* instead of *-i* or *-y*, especially in the final position; the usage of *-ee*, when a word finishes in *-e*; the addition of *h* to express attitude and accent; the replacement of *o* by *w* in any position of the word; replacement of *-o* by *-oe* for the diphthong /əʊ/; and the usage of *-sz* for plurals. Further research should be carried out to check the reasons for its usage, but the reasons for the appearance of this new device might be: (a) to show emphasis in pronunciation/intonation/accents; (b) enthusiasm; (c) to show attitude; (d) to show that users are fashionable; and (e) to produce a message that tries to reflect slang language. The significance of this new device is that SMS language seems to have evolved into a fashionable and stylish way of writing where shortened versions of the words are not always the aim of the respelling. Texters seem to prefer to lengthen some words instead of typing the word without any spelling adaptation or a shortened version of the word.

Moreover, in MOC there are a few instances of number 3 being used as letter *-e*. Obviously, those examples were treated as non-conventional spellings rather than as letter/number homophones, as the number 3 is not pronounced as such but as the pronunciation corresponding to the letter *-e* depending on the context. There were no examples of this usage in the OC. This also supports the idea that SMS language seems to have evolved into a fashionable and stylish way of writing where the way of writing is as important as the content. It also seems that shortened versions of the words are not always the aim of the respelling.

2.4 Other significant features

After a deep analysis of the pupils' transcriptions (MOC) it has come to my notice that a significant number of instances of slang words and colloquial expressions were employed throughout the text messages. As those words have not been altered orthographically, they have not been classified in the devices mentioned before, notwithstanding the fact that these words attempt to imitate the way people speak. The following cases serve as an illustration: *yo* ('hello'), *hiya* ('hello'), *hol-lah* ('hello'), *skeen* ('I see'), *peng* ('sexy/hot'), *boomting* ('very good looking'), *buff*

(‘good looking’ in the context of this study, only used in London), *crap* (‘rubbish’), *mandem* (‘friends/person/men’), *niggah* (‘friends/black’, not a racist word), *blad* (‘brother’, comes from the Jamaican pronunciation of blood), *copp* (‘buy/win’), *safe* (‘thanks’), *don’t gas* (‘don’t lie’), *gwarn* (‘going on’), *bare cute* (‘very cute’), adding the word *man* to the sentences, replacing the 1st singular personal pronoun (‘I’) by *me*, the use of the ungrammatical expression *ain’t* (‘I have not/I am not’), the ungrammatical use of the verb *to be* with *you* (‘how is you’), incorrect usage of auxiliaries with *he* or *she*, such as *he don’t*, or the usage of *innit* for a question tag. In the OC there is not such a frequent usage of slang and colloquial expressions. This suggests that there is an increasing tendency to use this type of language in text messages.

3. Conclusions

There is a set of devices employed when sending text messages: reductions (shortenings, contractions, g-clippings and other types of clippings, abbreviations, acronyms, initialisms, ‘misspellings’ and typos), phoneticised respellings (letter/number homophones, non-conventional spellings and accent stylisation), word deletion and punctuation (lack of function words, lack of punctuation, over-punctuation and capital letters), smileys and other symbols. The use of these devices has been popular and extensively used in SMSes throughout the first decade of the 21st century, but it is presumably in order to avoid ambiguity and increase speed that the adaptations based on the sound-to-spelling relationship, *i.e.*, phonological approximation, are the most frequent ones, as the data suggest. This supports the idea that there is an intrinsic relationship between English pronunciation and spelling regardless of the independence that could be expected at first sight, as native speakers are able to respell words in such a way that the word is shorter but it is pronounced in exactly the same way as the original word. Moreover, native speakers are able to understand phoneticised respellings even when other texters use them for the first time because those words have the same pronunciation. Thus, native speakers establish a relationship between English spelling and pronunciation naturally.

The data obtained in MOC suggest that SMS language seems to have evolved into a fashionable and stylish way of writing, where shortened versions of the words are not always the aim of the respelling. The most significant change is the use of ‘stylish talk,’ a new device which consists in lengthening words to emphasize accent, slang, and attitude. The reasons for the appearance of this new device might be: (a) to show emphasis in pronunciation/intonation/accent; (b) enthusiasm; (c) to show attitude; (d) to show that users are fashionable; and (e) to produce

a message that tries to reflect slang language. Brevity does not seem to be the aim of SMS language anymore. Representing the way people talk when writing seems to be the new tendency in SMS language. In addition, the use of slang and ungrammatical expressions has increased. However, it should be borne in mind that there are significant differences in the two corpora such as the number of SMS analysed (202 in the OC and 1,500 in MOC) or the participants' age (17–18 year-olds and 12–14 year-olds respectively). Thus, further research should be carried out to ensure that independently of the age range (children, teenagers, young adults or adults), phoneticised respellings are the most recurrent devices in SMS language. Further research is required to check that the use of 'stylish talk' is not a regional device from London.

References

- Androutsopoulos, Jannis. 2000. Non-standard spellings in media texts: The case of German fan-zines. *Journal of Sociolinguistics* 4(4): 514–533. Available at: http://jannisandroutsopoulos.files.wordpress.com/2009/12/jos_4-4_fanzines.pdf
- Baron, Naomi S. 1998. Letters by phone or speech by other means: The linguistics of email. *Language and Communication* 18: 133–170.
- Cannon, Garland. 1989. Abbreviations and acronyms in English word-formation. *American Speech* 64/2: 99–127.
- Crystal, David. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.
- Crystal, David. 2003. TXT NE1? In *The Dyslexia Handbook* (Reading: British Dyslexia Association, 2006, S. Tresman, A. Cooke (eds.)), 179–83, [expansion of New Statesman 2003 article].
- Crystal, David. 2004. *A Glossary of Netspeak and Textspeak*. Edinburgh: Edinburgh University Press.
- Crystal, David. 2008a. The joy of txt. *Spotlight*, November 2008: 16–21.
- Crystal, David. 2008b. Texting. *ELT Journal* 62(1): 77–83.
- Crystal, David. 2009. *Txting: The Gr8 Db8*. Oxford: Oxford University Press.
- Dean Brown, James. 2009. Open-Response Items in Questionnaires in Qualitative Research in *Applied Linguistics*. In *A Practical Introduction*, J. Heigham & R. Broker (eds.), 200–216, Basingstoke: Palgrave Macmillan.
- Frehner, Carmen. 2008. *Email — SMS — MMS. The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern: Peter Lang.
- Grinter, Rebecca E.; Margery A. Eldridge. 2001. y do tngrs luv 2 txt msg? In *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work*, W. Prinz, M. Jarke, Y. Rogers, K. Schmidt; V. Wulf (eds.), 219–238, Netherlands: Kluwer Academic Publishers. Available at: <http://www.cc.gatech.edu/~beki/c14.pdf>
- Kunda, Sue. 2005. Teens Let Their Fingers Do the Talking. Available at: <http://www.txt2nite.com/forum/viewtopic.php?t=246>
- López Rúa, Paula. 2002. On the Structure of Acronyms and Neighbouring Categories: A Prototype-Based Account. *English Language and Linguistics* 6/1: 31–60.

- López Rúa, Paula. 2004. Acronyms & Co.: A typology of typologies. *Estudios Ingleses de la Universidad Complutense* 12: 109–129.
- López Rúa, Paula. 2007. Teaching L2 vocabulary through SMS language: Some didactic guidelines. *ELIA* 7: 165–188.
- Markus, Manfred. 2010. Spokenness as a feature of Late Modern English (with a focus on the 19th Century). Unpublished plenary session at the *Second international conference on Corpus Linguistics*, A Coruña University, A Coruña, Spain.
- Online SMS corpus which used to be available at: <http://www.netting-it.com/txt8.html>
- Shortis, Tim. 2001. *The Language of ICT — Information and Communication Technology*. London/New York: Routledge.
- Shortis, Tim. 2007. Gr8 txtpeceptions: The creativity of text spelling. *English Drama Media* 8, NATE: Sheffield, 21–26.
- Thurlow, Crispin. 2003. Generation Txt? The sociolinguistics of young people's text messaging. *Discourse Analysis Online (DAOL)*. Available at: <http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003.html>
- Thurlow, Crispin; Susan McKay. 2003. Profiling 'new' communication technologies in adolescence. *Journal of Language and Social Psychology* 22(1): 94–103.

Summary

SMS language is regarded as a 'new' communication system (cf. T. Shortis 2007; D. Crystal 2009; M. Markus 2010) characterized by new relationships that native speakers establish between English spelling and pronunciation by using different respelling devices (cf. C. Thurlow 2003; P. López Rúa 2007; D. Crystal 2009). This paper is an attempt to contribute to this recent area of study by analysing the development of SMS language over the last 10 years. Recent findings suggest that even though SMS language might have emerged out of the need for speed and brevity (C. Thurlow 2003: 4) — every SMS has a limited amount of characters —, it seems to have evolved into a fashionable and stylish way of writing where shortened versions of the words are not always the aim of the respelling. For the purpose of a diachronic analysis a free online British SMS corpus available at [netting-it.com](http://www.netting-it.com) containing 201 text messages compiled in 2000 has been used in comparison with my own data obtained in May 2010 by means of questionnaires carried out in a London secondary school. Thus, it was possible to analyse the differences in the use of SMS language during the last decade. The research proves that one of the most significant changes is the use of 'stylish talk', a new device which consists in lengthening words to emphasize accent, slang, and attitude. This contrasts with the general belief that words are merely shortened in SMS language. Moreover, the use of slang and ungrammatical expressions also seem to be frequent devices.

Text4Science

A Quebec French database of annotated text messages

Philippe Langlais and Patrick Drouin
University of Montreal

Introduction

Short Message Services are used by a huge community of users for an increasing number of purposes, including advertising (F. Bamba and S.J. Barnes 2007, R. Wei *et al.* 2010), voting (for instance for TV programmes), and even political campaigning (D. Ahvazi 2004). According to Wikipedia, more than 4 trillion text messages (or SMSes) were exchanged in 2008. The International Telecommunication Union¹ indicates that this figure roughly tripled from 2007 to 2010 to a staggering 6.1 trillion text messages, while the number of Internet users ‘only’ doubled between 2005 and 2010.

The ubiquitous nature of text messaging has inspired a large number of studies focusing on various aspects, including social ones (D. Reid and F. Reid 2004, L. Leung 2006, J. Wajcman *et al.* 2007, N.S. Baron 2008), linguistic ones, e.g. (J. Anis 2001, L.-A. Cougnon 2010) as well as educational ones (E. Scornavacca *et al.* 2007, S. So 2009). It is also natural to observe a growing interest in technological aspects related to text messages. Text completion for traditional touchtone phone keypads has been among the first applications studied (S. McKenzie *et al.* 2001, How 2004). Since then, the development of smart keyboards designed to ease the entering of text on smartphones and other portable devices (e.g., tablets, game consoles, etc.) has evolved quite drastically. The Swipe application² as well as the Swiftkey keyboard³ are two striking illustrations of how fast the mobile phone technology is evolving.

1. <http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf>

2. <http://www.swype.com/>

3. <https://play.google.com/store/apps/details?id=com.touchtype.swiftkey&hl=en>

One of the most recently studied applications is SMS normalisation, that is, the transformation of SMS-like texts into their ‘standardized’ version. This has been studied for instance for the English language by (A. Aw *et al.* 2006, M. Choudhury *et al.* 2007), as well as for the French language, e.g., (F. Yvon 2010 R. Beaufort *et al.* 2010).

Perhaps (or hopefully) more marginally, technologies are deployed in a car environment, either for reading SMSes by a text-to-speech synthesis system, as in the Ford Sync system, or to assist a driver to answer a message while driving (Y.-C. Ju and T. Paek 2010). Also, Munro (2010) describes the service deployed by a consortium of volunteer organizations named “Mission 4636” during the earthquake that struck Haiti in January 2010. This service routed SMSes alerts reporting trapped people and other emergencies to a set of volunteers who translated Haitian Creole SMSes into English, so that primary emergency contributors could understand them. In (W. Lewis 2010), the authors describe how, in the same context, the Microsoft Translation team developed a statistical translation engine (Haitian Creole into English) in as few as 5 days.

As noted by C. Fairon *et al.* (2006), most of the aforementioned studies could only be conducted thanks to the availability of corpora of SMSes in different languages. Such corpora are available in some languages. For instance, the NUS SMS corpus consists of 10,117 English SMSes collected from students of Singapore University who were asked to type (over a web form), messages they received or sent. Some messages acquired from chats complement the collection. A live version of this corpus is also available online (T. Chen and M.-Y. Kan 2012). In October 2011, 28,724 English SMSes were gathered by 116 contributors and 29,100 Chinese ones by 515 contributors. Also, the British English SMS Corpora⁴ gathered slightly more than 800 SMSes. In M. Treurniet *et al.* (2012), the authors describe the SoNaR SMS corpus. They managed to acquire about 53,000 Dutch messages over the rather short period of three months. A good overview of available corpora and their main characteristics can be found in T. Chen and M.-Y. Kan (2012).

A number of corpora are also available for the French language (some are proprietary); the largest collection coming from SMS4science, an international project coordinated by the Université Catholique de Louvain in Belgium. This international project gathers corpora in various languages by asking participants to forward SMSes they already sent. This way of acquiring text messages avoids typos introduced by copying text messages over a web form, and guarantees to some extent that the messages sent are real ones. French corpora collected in different regions or countries are already available: Belgium (*sms4science*), Reunion Island

4. <http://mtaufiqnzz.wordpress.com/british-english-sms-corpora/>

(*LaRéunion4science*), Switzerland (*sms4science*) and France (*smsAlpins* as well as *sud4science*). We refer the interested reader to the website of the SMS4science's project for an extensive list of articles related to SMS.

This article describes the Quebec French corpus we collected within the SMS4science project, namely the textto4science corpus. In the remainder, we describe in Section 1 the design of the textto4science project. We present in Section 2 the rich annotation system we adopted for preparing the SMSes we received. We provide a linguistic and a sociolinguistic account of our database in Section 3 and 4 respectively. We discuss this study and a number of ongoing projects in Section 5.

1. Design of the Textto4science project

In October 2009, the Quebec French part of the international SMS4science project called textto4science was launched. Over a period of 10 months, we collected slightly more than 7,000 SMSes that we carefully annotated. This collection and annotation phase was made possible thanks to a grant we received from the Centre for Multidisciplinary Research on Emerging Technologies (Centre de recherche Interdisciplinaire sur les Technologies Émergentes — CITÉ)⁵ from the University of Montreal. This grant was spent mostly on renting a short number to which donators could send previously written SMSes, and to pay the two students who conducted the annotation. A small amount of money was also devoted to advertising the project.

1.1 Technical platform

Early in the project, we decided to rent a short number (202202) to which donators could resend SMSes stored in the cache of their telephone device. We preferred this scenario to a dedicated webpage where people could type in SMSes. This option avoids errors introduced by users when typing their messages over the interface, and guarantees to some extent that the messages they submit were not written specifically for the data collection, but were indeed messages they already sent to someone. Our industrial partner *Adenyo Télécom Mobile Inc.*⁶ kindly provided us with a technical platform to facilitate the management of the messages we received on the short number.

5. <http://www.cite.umontreal.ca/>

6. <http://www.adenyo.com/>

1.2 Elaboration of an annotation policy

The annotation policy we designed in this project is the result of an iterative procedure involving two students we recruited at the linguistic department of the University of Montreal for annotating the SMSes we received. The annotation took place in roughly four stages. In a first one, they annotated a number of messages, using plain text annotation, typed with Excel as an interface, and without many instructions regarding the way messages should be annotated. This stage served to observe the various phenomena encountered in SMSes. In a second stage, we documented those phenomena (P. Drouin et al. 2010) and adopted XML for annotating our SMSes. Based on this document, we developed an XML schema that was used within the user-friendly XML editor <oxygen/>.⁷ In a third stage, we refined the XML system used for annotating, and the annotators revised the annotation they produced. In a fourth and last step, one of the two annotators checked all the annotations with the goal of reinforcing their consistency.⁸

1.3 A rich annotation system

The main originality of our project lies in the richness of the annotation policy we devised. In order to appreciate this, it is worth looking at Figure 1, which reports an excerpt of the database, gathered within the Belgian SMS4science. The database takes the form of an Excel spreadsheet, where one column stands for the original message and the other stands for the normalisation or transcription. While most words are normalised, some are not. For instance, the form *bizo* is marked as ambiguous and is not normalised. This form likely stands for *bisous* (literally, kisses) or one of its SMS variants, such as *bisouX*. Also, the form *jtm*, which very likely stands for *je t'aime* (literally, I love you) is not normalised either.

original: *Ccou princes jpens trop a ti et tme mank tro.. Jesper kon svoi bi l tot..
Gro gro bizo jtm*
annotation: *Coucou princesse je pense trop à toi et tu me manques trop..
J'espère qu'on se voit bientôt.. Gros gros {bizo,,AMBIG} jtm*

Figure 1. Example of an annotated SMS in the Belgian SMS4science corpus. AMBIG indicates that the form *bizo* is ambiguous

Our annotation scheme differs from the one just pictured in several aspects. First, we decided to produce an XML database. This allows a clean separation of the text

7. <http://www.oxygenxml.com/>

8. As we discuss in Section 5.1, we failed to track a few important matters, as our main goal was to make the corpus available. In particular, we did not keep a record of the number of cases where both annotators initially disagreed.

data from the annotation, which is not the case of the format shown in Figure 1. It also facilitates the alignment of the normalised form with the original SMS. Last but not least, it allows the validation of the database against an XML schema (or a DTD), as well as the manipulation of the database with XSL transformations. Those last two points proved to be very useful when came the time for checking and anonymising the database.

Second, we extended the number of linguistic phenomena typically annotated in similar databases with the aim of easing the normalisation of SMSes into proper Quebec French. On the downside, we must admit that the annotation task we defined is much more demanding than what has been done in other collections. The many examples we provide in this article witness that annotating SMSes at a fine-grained level is indeed difficult.

```
<texto ID="2009120927" date="2009-12-09">
  <user_id>user_111250</user_id>
  <orig>
    Salut Florent, je sais pas ou tu es, mais si tu peux
    connecte toi sur Skype, j y serai une partie de la
    soiree! Ciao
  </orig>
  <transcrip>
    Salut <prenom sexe="masc">Florent</prenom>, je
    <negat forme="ne"/> sais pas <ortho forme="où">
    ou</ortho> tu es, mais si tu peux<ponc forme=","/>
    <typog forme="connecte-toi">connecte toi</typog>
    sur Skype, <typog forme="j'y">j y</typog> serai
    une partie de la <ortho forme="soirée">soiree
    </ortho>! Ciao<ponc forme="."/>
  </transcrip>
  <norm>
    Salut Florent, je ne sais pas où tu es, mais si tu
    peux, connecte-toi sur Skype, j'y serai une partie
    de la soirée! Ciao.
  </norm>
</texto>
```

Figure 2. Example of an annotated and anonymised SMS. We slightly edited this XML instance for the sake of readability

Figure 2 provides an example of an SMS in the *texto4science* database. Each message (element *texto*) has a specific identifier (attribute name *ID*), and is stamped with the date at which it was received (attribute *date*). The donor identifier (element *user_id*) indicates which person sent this message; it can be used to consult the sociolinguistic information provided by the person (if any, see Section 4). The text message received is encoded in the *orig* element, after anonymisation took

place (replacement of names, phone numbers and the like with generic entities). The *transcrip* element contains the annotation of the original message. There are basically two families of elements that are used to annotate an SMS: those that denote specific units (such as a laugh indicator), and those that denote errors (such as a typo), in which case the *forme* attribute contains the corrected text. Each element is detailed in Section 2. Last, the *norm* element contains the normalisation of the SMS. It has been automatically produced by applying XSL transformations to the *transcrip* element, another benefit of using XML technology.

We made a design choice which turned out to be a limitation of our annotation policy:⁹ we decided to use non-recursive XML elements in our annotation system. This means that only one element can be associated to a given portion of a text message. For instance, an abbreviation in English such as *asap* (as soon as possible) is annotated as being an English block of text (element *bloc_lang*) in our database, but not as an abbreviation. Sometimes, the attribute *comment* is used to document such situations.

1.4 Diary of the Texto4science project

Before we started receiving the first SMS, we experienced a number of difficulties whose details are reported in P. Langlais *et al.* (2012). The main difficulty was getting the project accepted by the ethical committee of the University of Montreal. We were also disappointed by our failure to obtain from the Canadian Wireless Telecommunications Association¹⁰ a removal of the fees incurred to calls to the short number we rented for the project. This had a significant impact on the number of SMSes we received in the end.

A great deal of effort was spent on promoting the project in various public media, including radio and TV programmes and major social networks. Figure 3 shows the evolution of the number of SMSes received over time. Retrospectively, two events had a strong impact on the number of messages we received. The first one was on 26 November 2009 when Patrick Drouin was invited to Radio Canada for a popular radio programme called “Christiane Charette”.¹¹ He was accompanied by Biz, the singer of the hip-hop band *Locolocass*,¹² a famous band in Quebec which promotes French in Quebec. This event corresponds to the first peak in

9. We could not afford the time for re-annotating those cases.

10. <http://www.cwta.ca/>

11. http://www.radio-canada.ca/emissions/christiane_charette/2009-2010/chronique.asp?id-Chronique=97193

12. http://fr.wikipedia.org/wiki/Loco_Locass

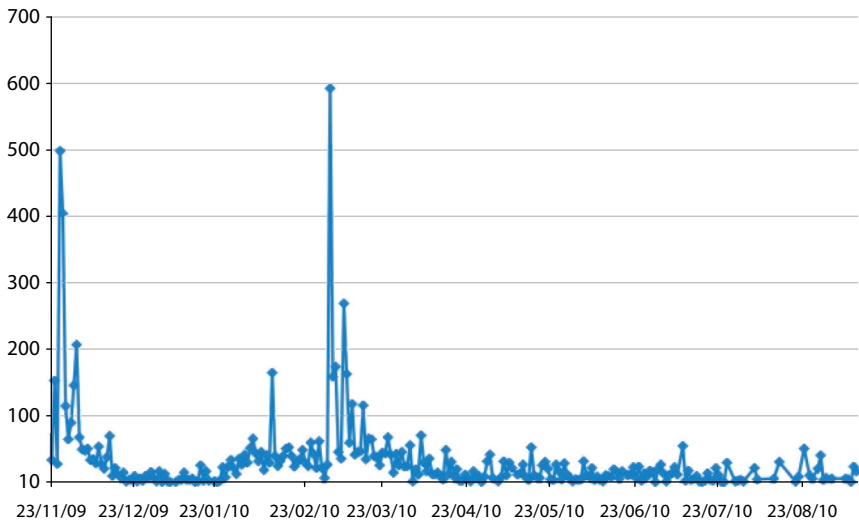


Figure 3. Number of SMSes received over time

the number of messages we received over time, as shown in Figure 3. The second peak we observed was on 4 March 2010 when Patrick Drouin was interviewed on a popular TV programme called *Salut, Bonjour*.

All in all, we processed a total of 7,274 text messages, sent from 360 different phone numbers. Among those messages, we received 420 (5.8%) SMSes written in English, 6 written in Spanish, 1 written in Italian and 5 in other languages. Those SMSes did not receive any annotation apart from the fact that they are written in a language other than French. Similarly to other SMS collection projects, a few donators contributed a lot to the database. As a matter of fact, we noted that half of the SMSes we gathered came from 11 contributors only, while the main contributor (a female) provided us with 1,100 SMSes (15%).

2. Annotation system

We used two families of elements for annotating an SMS: those that denote specific units present in SMSes, such as smileys, and those that denote errors, such as typos, for which the *forme* attribute contains the corrected text. Those two types of elements are described in Section 2.1. and 2.2. respectively.

Element	Count	Excerpts of SMSes	Attributes
abrev	5082	Keskon fait ...	forme=Qu'est-ce qu'on
binet	1605	...ta toune a la radiooo xD	
bloc_lang	1480	Watch out, j'arrive !	langue=en
prenom	1038	...bienvenu pour la game Claude ...	sexe=andro
rire	622	AAA HAHAAHAAAA! !!! ...	
symb	183	De Coco @ Cocotte ...	forme=à
forme_inc.	112	envoyer que 2 mtn d message ...	
nom	112	Happy birthday Cottin! :)	
element_in	28	Maudit auto-correcteur ?????	
numero	21	123 456-7890. Ça c'est son ...	
adresse	17	444 Berrichon! Jpas partie ...	
mail	1	...luh.. berrichon.be@berriche.com	
web	1	pi tu met "http://berrichon.com" dans ...	

Figure 4. Representative examples of each of the 13 elements characterizing specific units annotated in our database. They are listed in decreasing order of frequency (count) in our database. The text being annotated is shown in bold. Some annotation elements are qualified by one or several attributes that are reported in the fourth column

2.1 Specific units

We annotated thirteen types of units we encountered in SMSes. Each one is illustrated in Figure 4 along with their frequency in our database. Abbreviations are one of the main characteristics of SMSes, and we analyse them in some detail in Section 3.2. We marked each abbreviation we found along with its plain form (element *abrev*). Similarly, we annotated symbols (element *symb*) that often serve as abbreviations (e.g., @ for the word *at*), but that can be used as punctuation marks as well. Smileys are typical of SMSes and are marked (element *binet*), as well as various laughing indicators (element *rire*), such as *ah ah* or *Mouahahah! !*. Also, parts of text messages that are in another language are marked by the *bloc_lang* element along with the language being recognized (attribute *langue*). In order to facilitate the anonymisation of the messages, we marked first names (element *pre-nom*), family names (element *nom*), numbers (element *numero*), e-mail addresses (element *mail*), URLs (element *web*), as well as postal addresses (element *adresse*). Last, two elements designate sequences of characters that we did not manage to understand. We distinguish cases where a sequence is composed of non-alphabetic characters (element *element_inc*) from others (element *forme_inc*).

Element	Count	Excerpts of SMSes	Attributes
ponc	9027	<i>Jte donne tu assez de detail⊗:)</i>	forme=?
ortho	7496	<i>Chéri la souffleuse c'est moi ...</i>	forme=souffleuse
synt	2924	<i>Est ce que au 31c tarrive a...</i>	forme=qu'au type=retab
		<i>Moi aussi! Me semble...</i>	forme=il type=ins
majus	1738	<i>...tu ten va loin genre st-calixte ...</i>	forme=St-Calixte type=ajout
		<i>L'heure, C'est parfait pour moi!</i>	forme=c'est type=supp
accord	1505	<i>Jespere tu tombera pas sur une ...</i>	forme=tomberas
		<i>...avait deja acheter 2 paquets ...</i>	forme=acheté
typog	1006	<i>...Virginie, je t,appelle plus tard.</i>	forme=t'appelle
negat	818	<i>donc on ne fait ⊗ les chemises ...</i>	forme=pas
		<i>...fie toi pas sur moi ...</i>	forme=ne te fie pas
			comment=synt corr
coquille	267	<i>...Bon courage ...</i>	forme=courage

Figure 5. Representative examples of the eight types of errors we annotated in our database. They are listed in decreasing order of frequency (count) in our database. The errors are illustrated in bold; a ⊗ sign marks indicates missing parts. Some annotation types are qualified by one or several attributes that are reported in the fourth column

2.2 Annotated errors

As well as describing typical units that are present in text messages, we also annotated some errors and their correction (via the attribute *forme*). We distinguished a total of eight errors that are illustrated in Figure 5 along with their frequency in our database.

In accordance with the Belgian database of the *SMS4science* project, we annotated missing negations that are frequent in spoken language, as well as incorrect constructions involved in a negation. Two negation errors are illustrated in Figure 5 (see element *negat*). We also annotated missing punctuations (element *ponc*) because SMSes are known to miss punctuation marks, which poses problems for SMS-to-speech converters, for instance.

Typos are marked by the element *coquille* (e.g., *trsiste* instead of *triste*), while typographical errors are tagged by the *typog* element (e.g., *repose toi* instead of *repose-toi*). Spelling errors are marked by the element *ortho*, as in *nous meme* for the form *nous-même* (ourselves). Syntactical errors — e.g., *idée cadeau* (idea gift), instead of *idée de cadeau* (idea for a gift) — are tagged with the *synt* element. Also, the *accord* element indicates agreement errors, such as *dit-lui* for *dis-lui* (tell him). We also annotated each error involving case (element *majus*). Most often, they concern the absence of a capitalized letter at the beginning of a message or a proper name.

2.3 A quantitative view of the annotations

Slightly more than 35,000 annotations are present in the current version of the database, that is, roughly five annotations per message. The SMSes we received are characterized by a high rate of missing punctuations (*ponc*), a large number of spelling errors (*ortho*) and a lot of abbreviations (*abrev*). 258 of the SMSes we received (3.8%) do not have any annotation, which means that they are written in standard Quebec French, without any noticeable error or SMS-like idioms (smileys, etc.). Although it is anecdotic, one long message¹³ received no less than 86 annotations.

3. Linguistic analysis

In this section, we describe the linguistic analysis we conducted over the 6,842 text messages we annotated and that were written in French (we must recall that 432 messages were written in other languages).

3.1 SMSes are slightly shorter than their normalised counterpart

It is often believed that SMS messaging is geared toward shorter texts, due in part to the length limitation applied by most telephone operators. In the corpus we gathered, we observed that SMSes are about 10% shorter than their normalised counterpart (compression rate). The average SMS length is 58 characters, while their transcriptions are 65 characters long on average. The compression rate was evaluated both at the character level and at the word level. Compression at the former level is interesting since SMSes have a theoretical limit measure in characters

Table 1. Main characteristics of the 6,842 SMSes written in French. A simple set of regular expressions has been applied in order to tokenize the material

	SMS	Normalised
Number of tokens	90 298	104 268
Number of types	11 750	9 279
Number of hapax	7 215	5 401
	Char	Word
Compression rate	10.5%	13.4%

13. Whilst the limit imposed by the SMS protocol is 160 characters, some operators authorize the manipulation of longer messages. The one we discuss is 563 characters long.

and not in words. Users can of course write messages that are longer than the limit enforced by the operator, but the messages are then split into multiple messages.

According to (A. Kornay 2007) we can estimate the ratio of hapax/type at 40% to 60% for English. Based on the fact that SMSes often contain tokens and abbreviations which are often not present in a more standard register, one would have expected this ratio to be higher than usual. Yet when we perform this calculation, we measure a 61.4% ratio in our corpus. The fact the hapax ratio obtained is fairly standard could be attributed to two different reasons: either the number of SMS-specific types is not as high as expected or the number of types itself is fairly high. This last explanation would allow us to conclude that people who write SMSes have good grasp of the French lexicon.

3.2 Abbreviations are numerous and ambiguous

One possible use of an annotated corpus such as ours consists in compiling a dictionary dedicated to SMSes. Some are already available, such as *www.dictionnaire-sms.com*, *www.sos-sms.net* or *www.deblok.net/dicosms*, but are nevertheless never large enough to account for the great creativity in SMSes. L.-A. Cougnon and R. Beaufort (2010) present a methodology to semi-automatically build up a dictionary out of an SMS corpus. Since we adopted an XML-based annotation system, building such a dictionary from our database is simply a matter of querying *abrev* elements and using an XSL transformation. 5,082 abbreviations have been annotated in our database, for a total of 1,256 different abbreviations. Those abbreviations concern a total of 1,047 normalised forms.

Figure 6 reports the ten most frequent forms abbreviated in our database, as well as their three most common abbreviations. We observed that some

<i>okay</i> (310)	→	<i>ok</i> (251)	<i>k</i> (56)	<i>okk</i> (3)
<i>c'est</i> (283)	→	<i>c</i> (270)	<i>c'</i> (5)	<i>cé</i> (5)
<i>dans</i> (174)	→	<i>ds</i> (168)	<i>dan</i> (4)	<i>dn</i> (2)
<i>que</i> (141)	→	<i>ke</i> (128)	<i>q</i> (13)	
<i>pour</i> (120)	→	<i>pr</i> (120)		
<i>avec</i> (92)	→	<i>ak</i> (48)	<i>aek</i> (22)	<i>ac</i> (11)
<i>tu es</i> (88)	→	<i>t</i> (69)	<i>te</i> (6)	<i>té</i> (5)
<i>parce que</i> (82)	→	<i>pcq</i> (51)	<i>pck</i> (20)	<i>paske</i> (6)
<i>il</i> (71)	→	<i>y</i> (70)	<i>ii</i> (1)	
<i>vous</i> (71)	→	<i>vs</i> (70)	<i>vou</i> (1)	

Figure 6. The 10 most frequently abbreviated forms and their (at most) three most frequent abbreviations

abbreviations are ambiguous, such as *dispo* which stands for both *disponibilités* (availability) and *disponible* (available). The same can be said of the form *l* which stands both for *une* (a, feminine) and *un* (a, masculine). The most ambiguous abbreviation we found is *txt* which is used for various morphologically derived forms of the word *texte* (*text*): *texte*, *texteras*, *texté*, *textes*, *texter*, *texto*.

It is interesting to note that while some abbreviations are purely lexical, such as *Prof* for *professeur* (professor), others involve compounds, such as *fds* for *fin de semaine* (weekend) or sometimes phrases, such as *tlm* for *tout le monde* (everybody). Abbreviations such as these are often well integrated and known by the general public and are sometimes very specific to SMS-like language, such as *pcq* (because).

We also noticed a great variability in common expressions, such as *à plus tard* (see you later) which we found written as *a+* (9), *aplus* (2), *a pluche* (1), *a plus* (1), *aplush* (1) and *à plus* (1), where the figures in brackets indicate the frequency of the form in the database. Similarly, we found six different ways of writing the word *demain* (tomorrow): *2m* (3), *dmain* (2), *2mai* (1),¹⁴ *2main* (1), *2min* (1), and *dmin* (1). Fairon *et al.* (2006) also noticed a large number of variants for this word.

3.3 On the use of smileys

As Figure 4 shows, 1,605 smileys (element *binet*) have been annotated. We identified a total of 98 different smileys in our corpus. This is much less than the 900 observed in Beaufort *et al.* (2010). The three most frequent ones are :) (37.3%), :P (6.8%) and :((6.2%). Some less common smileys are nevertheless creative such as >:-> (which we observed three times), or (>_<) which we observed only once. In the same vein, we annotated 622 tokens that refer to laughing (element *rire*), for a total of 113 different forms; many being variants of the same form (e.g., *hahahaha*, *hahah*), others being more surprising (e.g., *L0olehw'zz*).

3.4 Missing annotations

Despite the richness of our annotation policy, there are phenomena that we did not have time to analyse. In particular, we did not mark expressions typical of Quebec French, since we concentrated mainly on deviations in SMSes from standard Quebec French. Those expressions are still numerous in our corpus and deserve dedicated analyses. Some are English calques, such as *prendre une marche* (take a walk), or *pour vrai* (for real). Others involve words that do not exist in French, such as *magasinage* (shopping) for which a native French speaker would likely use the English form.

14. Which is likely to be a typo.

Over the 6,842 SMSes written in French which we annotated, we identified no less than 1,480 parts (words or phrases) written in another language. While we marked those expressions as foreign ones, we did not annotate them more precisely. Some examples involve loan words (e.g., *worker*, *casual*, *news*) while others contain phrases (e.g., *see you soon*). Sometimes certain English words are also written in an SMS way (e.g., *2day* for today). Also, some cases involve a kind of transliteration process, such as *je bouque* (I book).

The fine-grained annotation of both Quebec French expressions and code-switching in our database is an ongoing effort which is discussed in Section 5.2.

4. Sociolinguistic analysis

Volunteers who gave their SMSes to the texto4science project were invited to fill in a survey containing 23 questions designed to collect their profile. A total of 298 persons responded. The questions are grouped into four main categories: personal information (age, gender, postcode, mother tongues, spoken languages and education), use of SMSes (average number of SMSes sent per week, usual places where messages are sent and received, categories of persons to whom most messages are addressed, etc.), abilities in writing SMSes (participants' familiarity with abbreviations and other codes frequent in text messages; their use of such idioms in their production; the fact that they mix several languages in a single SMS, etc.) and the technical device used for texting (12-touch regular keypad, QWERTY keyboard, touch screens, etc.).

4.1 Contributors' profile

4.1.1 Personal information

Among the 298 respondents to the survey (63% females), 209 provided a telephone number corresponding to one that matches SMSes of our database. Those 209 contributors represent 58% of the persons (that is, telephone numbers) who gave their SMSes to the project.

The average age of the participants is 27, the youngest person is 12 years-old, the oldest 65. We often associate SMSes with young people. This is also reflected in our data since over 72% of our participants are aged between 12 and 39. See the age distribution in Table 2 for more details.

Slightly less than 30% (89) of the participants said they were students. Considering that we spent a lot of effort to promoting the project among colleges and universities, this somehow low rate comes as a surprise. Nevertheless, 60% (177) of the participants received a university education — 45 of them at graduate

Table 2. Age distribution of the 298 respondents

Age	# Contributors	%
12–19	57	19.1
20–29	160	53.7
30–39	42	14.1
40–49	28	9.4
50–59	9	3.0
60–65	2	0.7

level — and 26% (77) went to college. These figures indicate that our contributors are mainly educated people.

Most contributors (284) have French as their mother tongue, which is not surprising since we targeted SMSes written in Quebec French; 75% (224) of persons also mentioned they speak English fluently. Looking at their postcodes, we noticed that most of our participants are located in Montreal, and that a significant part of them are located downtown. A few of our participants live elsewhere in Canada. The concentration of our participants in the Montreal region underlines the difficulty we had in motivating people all over Quebec (and across the country) to donate their SMSes.

Table 3. Weekly time spent reading and writing

	≤1h	≤2h	≤3h	≤4h	≤5h	≤10h
<i>Reading</i>	19.8%	47.3%	63.1%	74.8%	84.5%	98.6%
<i>Writing</i>	57.1%	77.9%	90.0%	94.4%	96.4%	99.7%

We also asked the respondents about their reading and writing habits. Roughly half (47.3%) of our respondents answered they read two hours or less a day (Table 3). Most of them read content on the Internet (79.5%), but participants also read traditional books (73.8%) (see Table 4).

Table 4. Type of material being read

Press	Comics	Web	Song lyrics	Books	Others
229	33	237	51	220	119

Our contributors seem to spend less time writing than reading. Only 22.5% (67) of the participants indicated that they write for “casual communication” while 56.4% (168) selected the “Other” answer and commented that taking notes (for courses, work, etc.) is their most frequent writing activity (see Table 5).

Table 5. Other communication methods used

Email	Chat	Social Networks	Forums
287	220	249	56

4.1.2 Use of SMSes

The way participants use the SMS technology is summarized in the next seven tables. We can see that most of the contributors to our study are quite familiar with texting. Only 6% (18) of them are new users (texting for less than six months), while 12% (36) of our respondents had been texting for at least four years.

Table 6. Number of months users have been using SMSes

Duration	≤6m	≤1y	≤2y	≤3y	≤4y	>4y
<i>Users</i>	18	23	65	53	39	36

In the survey, we asked people to evaluate the number of SMSes received (in) and sent (out) on a weekly basis,¹⁵ a significant portion of participants (47%) deal with a few SMSes a week only (less than five messages received and sent). Only five respondents were sending (and receiving) more than one hundred messages a week. Therefore, it does not seem that we have cases of compulsive texters in our database.

Table 7. Number of SMSes sent on a weekly basis

Number	<5	≤10	≤20	≤50	≤100	>100
<i>Received</i>	135	63	46	36	13	5
<i>Sent</i>	40	58	48	32	15	5

The three last questions (Please rank the intended recipients of your text messages in descending order; Why do you use text messaging? and Where do you use text messaging?) detailed in Tables 8, 9 and 10 were formulated as multiple-choice questions. Each responder could rank each option; a score of 1 being associated to the option the most appropriate and a score of 5 to the least appropriate one. Options which were not relevant were marked as such.

For each of those questions, our tables contain three lines: *Contributors* which indicates the number of respondents who marked a specific option as relevant, *Rank 1* which indicates the number of participants who gave an option the rank of 1, and *Average rank* which indicates the average rank per option (not counting the ones that were judged irrelevant).¹⁶

15. Obviously both figures are highly correlated.

16. An average rank close to 1 indicates a very appropriate option.

Table 8. Main recipients of SMSes sent by out participants

	Family	Friends	Partner(s)	Colleagues	Games	Other
Contributors	250	291	201	208	79	91
Rank 1	38	151	111	12	2	4
Average rank	2.7	1.7	1.8	3.2	4.9	4.5

For instance, 291 respondents (that is, most of them) indicated that they send SMSes to their friends. Only 151 persons ranked this usage first (*Rank 1*), even if the average rank of this option is rather low (1.7). All those figures contribute to indicate that for the most part, participants send SMSes to their friends or partners, which is not entirely surprising.

Table 9. Main reasons for using SMSes

	Tel./e-mail	Cost	Inf. exch.	Meeting	Keep in touch	Chat
Contributors	279	223	275	243	253	242
Rank 1	145	58	55	36	47	56
Average rank	2.2	3.3	2.7	3.3	3.1	3.3

When asked about their motivations for using the SMS technology, most participants indicated that they use it mainly as a replacement for e-mails and telephone calls (*Tel./email*). 75% (223) of them indicated that reducing the cost of their mobile phone bill was a concern (*Cost*), although this was ranked 3.3 on average; only 19% mentioned this was their first motivation. Exchanging information (*Inf. exch.*), organising meetings (*Meeting*), keeping in touch with friends (*Keep in touch*) and chitchat (*Chat*) were options considered relevant by respondents.

People read or write SMSes in various places, such as at home, at work (or school), on transport and in public places. 42% (124) of them are doing so on a regular basis at work. One might assume that most people would rather prefer to write than call their friends, partner or spouse when at work (or at school) in order to avoid disturbing the other person in their professional or student life.

In addition to the previous data, we have collected 93 open comments made by participants that we have not analysed carefully. Often, our respondents mention that they are doing their best to write in proper (Quebec) French.

Table 10. Location where SMSes are written

	Home	Work/School	Public places	Transportation
Contributors	286	268	294	275
Rank 1	85	124	89	87
Average rank	2.5	2.1	2.1	2.4

4.1.3 Abilities to write SMSes

Only half of the participants (151) mentioned that it was easy for them to understand abbreviations typically encountered in SMSes. Code-switching is a common practice among contributors: 73.5% (219) mentioned they switch from one language to another from time to time. Although English is the language they switch to most frequently, other languages are being used as well, including Spanish, German, and Arabic.

4.1.4 Technical device

Technical aspects related to the way people write SMSes are summarized in the next two tables. First, it is noticeable that our respondents are not making a great use of dictionary facilities: only 16% (11+36) are making use of suggestions proposed by such tools. 17% (50) do not even know about this technology (those marked ??).

Table 11. Use of provided spelling dictionary

Never	??	Always	Often	Sometimes	Never
83	50	11	36	44	74

The way our contributors use dictionaries is certainly related to the kind of technical device they are using. In fact, 50% (148) use a standard 12-key keypad (*regul*), 31% (92) use a QWERTY keypad and only 15% (44) use touch screens. The kind of device used for texting is evolving fast, and the impact this evolution has on the quality of the SMSes produced deserves some investigations, for which our corpus will likely be useful. Several participants mentioned that typing accents on touch screens is difficult (often, it requires switching keyboards), and that typing with a QWERTY keyboard reduces the use of SMS-like idioms.

Table 12. Type of keyboard used

Standard	QWERTY	Touch screen	Pen	Other
148	92	44	3	11

Table 13 reports the average number of annotations per SMS. We observe that this figure is higher for regular telephones, and lower for users of touch screen phones. Also, the percentage of SMSes that did not receive any annotation (unannotated SMSes) is higher for touch screen users.

Table 13. SMS annotations with regards to the type of keyboard used

	QWERTY	Standard	Touch
Average # of annotations	4.6	5.0	3.2
% of unannotated SMSes	8.7%	9.2%	11.8%

Table 14 details the rate of an element annotation compared to the type of keyboard used. The rate of an annotation is computed by the ratio of the number of SMSes to the number of annotations. A rate of 1 would for instance indicate that a specific annotation is found in every SMS, while a rate of 10 would indicate that an annotation is found in one in 10 SMSes on average.

Table 14. Rate of annotations with regard to the type of keyboard used

	QWERTY	Standard	Touch
<i>abrev</i>	2.1	1.8	5.5
<i>accord</i>	7.4	5.7	7.9
<i>coquille</i>	34.4	33.4	32.9
<i>majus</i>	3.3	4.9	6.9
<i>ortho</i>	1.2	1.1	2.3
<i>ponc</i>	1.1	0.9	1.2
<i>synt</i>	3.8	2.8	6.9
<i>typog</i>	19.1	8.2	19.3

We observe that globally, the rate of errors (typos, syntax, absence of a negation, and typographical errors) and abbreviations is higher (closer to unity) for the users typing with a standard keypad. For instance, users of touch screen phones use abbreviations once every five messages on average, while users of standard keypads use them more than once every two messages. It is also interesting to note that the rate of laugh indicators (*rire*) and smileys (*binet*) is much lower among touch screen users (one laugh mark every 59 SMSes on average). One possible explanation for this is that touch screen phone users are using their device for professional communications. Obviously, a more careful crossing of the information collected during the survey with the SMSes we received needs to be accomplished to better appreciate those observations.

It should be noted that the last two tables were obtained by crossing the answers to the survey with the SMSes, therefore, it excludes the messages sent by persons that did not answer the survey or those sent from a telephone number not mentioned in the survey. Also, it is important to note that nothing prevent a responder from providing erroneous information, although from their comments, it looks like they were genuine and wanted to help the project.

5. Discussion

We have presented an overview of the texto4science project and its database which is freely available for download.¹⁷ Facilities for browsing online through the database are currently being refined, and will also be available. We hope this database will be helpful in developing applications targeting Quebec French, and in conducting comparative linguistic studies.

In the following sections, we list a number of lessons we learnt during this work and give a brief account of some ongoing projects.

5.1 Lessons learned

When we started this venture, we did not have a clear picture of the many activities that needed to be carried out for the project to evolve satisfactorily. We were naive enough to carry the project ourselves, without delegating much the work that needed to be accomplished. The lack of budget was actually the main reason for this approach.

Retrospectively, there are things that we would have done differently if we were to start the project again. First, and even if we were rather successful in doing so, advertising the project is a very demanding task. As noted by M. Treurniet *et al.* (2012) relying on the communication services of large institutions (such as universities) is a must. We did contact our communication department, but they were not entirely proactive, in part because we had barely any time to follow up on it.

Another thing we learnt is that each action conducted (from distributing flyers at a given place to simply posting a tweet) should be carefully logged. The analysis proposed by M. Treurniet *et al.* (2012) regarding the payoff of different promoting strategies is an example of invaluable information that could only be collected by carefully logging each promotional activity. We believe that there is a strong link between the number of incoming messages towards the last period of the project and our presence on social networks (Twitter and Facebook). The fact that we did not trace each action prevents us from measuring this precisely.

Last but not least, we definitely neglected the technical issues involved in collecting SMSes. The smartphone application *SMS Collection for Corpus* developed by T. Chen and M.-Y. Kan (2012) that automatically copies SMSes sent by a person to his/her *Gmail*¹⁸ account (so that it becomes easy to forward it to a dedicated email address) is a striking illustration that technical solutions that can greatly benefit the collection stage do exist.

17. <http://rali.iro.umontreal.ca/rali/?q=fr/texto4science>

18. <http://www.gmail.com>

5.2 Ongoing projects

We are currently working on several projects that we shall outline hereafter.

First, we are developing two translation engines, which are currently available online. The first one transforms a (Quebec) French text into an SMS-like text, such as *bonjour, jva bi1* for the message *bonjour, je vais bien*. This application makes use of lists we compiled from our database, notably abbreviations. The second engine normalises SMSes according to a statistical translation engine we trained on the *texto4science* database. This statistical engine is hybridized with rules that are designed to handle specific phenomena such as agglutinations, e.g., *Ca'c'peupa* (this cannot be true). Also, we observed in Table 9 that SMSes are often used for scheduling appointments. We therefore developed a strategy in order to recognize linguistic markers of appointments in SMSes and to extract relevant information (date, places, etc.). The prototype, called *CalTexto* automatically feeds Google Calendar with the extracted information (T. Bughin 2012).

On top of those applications, there are a number of issues regarding the present database we are currently addressing. First, we want to extend the markup language we used to annotate the SMSes in order to account for frequent phenomena we noticed while annotating and that deserve specific attention. Quebec French expressions are one phenomenon we have started to investigate. We also found code-switching to be very frequent in our database, another phenomenon we are currently looking at. So far 1,325 occurrences of code-switching have been identified in our corpus of approximately 7,000 messages. This number might seem rather high, but it reflects the bilingualism of Quebecers. A project in collaboration with the *Conseil supérieur de la langue française du Québec* is currently underway in order to describe code-switching between French and English among French speakers from the province of Quebec.

After the official end of the collection phase we conducted within the *texto4science* project, we received 7,174 SMSes that await annotation.¹⁹ We want to study whether we could annotate those messages semi-automatically, building on the annotations we have performed manually. This extra material will be a good addition to the database we described in this article.

Last but not least, we started the *text4science* project in collaboration with Simon Fraser University and the University of Ottawa. This project aims at collecting and analysing SMSes written in English across Canada. We just ended the data collection of this project and received over 15,000 messages that are currently being anonymised and annotated. For this project, based on the observation that

19. We did not close the short number to which donors continued to submit SMSes, probably because of our increased presence on social networks.

our rich annotation system incurred a high annotation cost, we opted for a less fined-grained XML annotation policy. We will, of course, make sure that this corpus and the one we described in this paper can be compared using a similar set of elements.²⁰ When ready, the new database will also be made available to the community.

Acknowledgments

We are very grateful to Amélie Paulus and Eugénie Rompré Brodeur the two annotators of our database without whom, nothing important could have been done. We also thank Florent Cottin who wrote most of the XSL transformations that we used for checking inconsistencies in the annotation, and for packaging the distributed version of our database. Fabrizio Gotti also kindly helped in distributing this database. We also thank the reviewers of this article for the useful comments and suggestions they made.

References

- Ahvazi, Daniel. 2006. Text messaging and its effects on everyday life. http://www.ahvazi.co.uk/pages/essays/Text_Messaging_and_its_Effects_on_Everyday_Life.html
- Anis, Jacques. 2001. *Parlez-vous texto? Guide des nouveaux langages du réseau*. Le Cherche Midi.
- Aw, Aiti; Min Zhang; Jian Xiao; Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *COLING/ACL on Main Conference Poster Sessions*, 33–40.
- Bamba, Fati; Stuart J. Barnes. 2007. SMS advertising, permission and the consumer: A study. *Business Process Management Journal* 13: 815–829.
- Baron, Naomie S. 2008. *Always on: Language in an Online and Mobile World*. Oxford: Oxford University Press.
- Beaufort, Richard; Sophie Roekhaut; Louise-Amélie Cougnon; Cédric Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *48th Annual Meeting of the Association for Computational Linguistics*, 770–779, Uppsala.
- Bughin, Thibault. 2012. *Analyse linguistique et extraction automatique de rendez-vous dans les SMS*. Master thesis, Université catholique de Louvain.
- Chen, Tao; Min-Yen Kan. 2012. Creating a live, public short message service corpus: The NUS SMS corpus. *Language Resources and Evaluation* abs/1112.2468.
- Choudhury, Monojit; Rahul Saraf; Vijit Jain; Sudeshna Sarkar; Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition* 10: 157–174.
- Cougnon, Louise-Amélie. 2010. Orthographe et langue dans les SMS. *Ela. Études de linguistique appliquée* 160: 397–410.

20. This implies that a less detailed annotation needs to be inferred from the current annotation of our Quebec French corpus, a task that should benefit the consistency of our annotations.

- Cougnon, Louise Amélie; Richard Beaufort. 2010. SSLD: A French SMS to Standard Language dDctionary. In *ELexicography in the 21st Century. New Challenges, New Applications*, S. Granger, M. Paquot (eds.), Cahiers du Cental 7, Louvain-la-Neuve: Presses Universitaires de Louvain.
- Drouin, Patrick; Eugénie Rompé Brodeur; Amélie Paulus. 2010. *Guide d'annotation*. unpublished report, University of Montreal.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier S. 2006. A translated corpus of 30000 French SMS. In *Language Resources and Evaluation*, 351–354, Genova.
- How, Yi Jue. 2004. *Analysis of SMS efficiency*. Honours Year Project Report.
- Ju, Yung Cheng; Tim Paek. 2010. Using speech to reply to SMS messages while driving: An in-car simulator user study. In *ACL 2010 Short Paper Session*, 313–317, Uppsala.
- Kornay, Andras. 2007. *Mathematical Linguistics*. Springer.
- Langlais, Philippe; Patrick Drouin; Amélie Paulus; Eugénie Rompré Brodeur; Florent Cottin. 2012. Texto4Science: a Quebec French Database of Annotated Short Text Messages. In *8th LREC*, Istanbul, Turkey.
- Leung, Louis. 2006. Unwillingness-to-communicate and college student's motives in SMS mobile messaging. *Telematics and Informatics* 24: 115–129.
- Lewis, William. 2010. Haitian creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAAMT*, Saint-Raphael.
- MacKenzie, Scott; Hedy Kober; Derek Smith; Terry Jones; Eugene Skepner E. 2001. Letterwise: Prefix-based disambiguation for mobile text input. In *ACM Symposium on User Interface Software and Technology*: 111–120, New York.
- Munro, Robert. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, Denver.
- Reid, Donna; Fraser Reid. 2004. *Insights into the Social and Psychological Effects of SMS Text Messaging*. University of Plymouth Working Paper.
- Scornavacca, E; S. Huff; S. Marshall. 2007. Developing a SMS-based classroom interaction system. In *Conference on Mobile Learning Technologies and Applications*, 47–54, Auckland.
- So, Simon. 2009. The development of a SMS-based teaching and learning system. *Journal of Education Technology Development and Exchange* 2(1): 113–124.
- Treurniet, Maaske; Orphe De Clerq; Henk van den Heuvel; Nelleke Oostdijk. 2012. Collecting a corpus of Dutch SMS. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC'12)*, 2260–2273, Istanbul, Turkey.
- Wajcman, Judy; Michael Bittman; Paul Jones; Lynne Johnstone; Jude Brown. 2007. *The Impact of the Mobile Phone on Work/Life Balance*. Rapport Interne, Australian Research Council.
- Wei, Ran; Hao Xiaoming; Ji Pan. 2010. Examining user behavioral response to SMS ads: Implications for the evolution of the mobile phone as a bona-fide medium. *Telematics and Informatics* 27(1): 32–41.
- Yvon, François. 2010. Rewriting the orthography of SMS messages. *Natural Language Engineering* 16(2): 133–159.

Summary

In October 2009 the Quebec French part of the international SMS4science project, called *text4science* was launched. Over a period of 10 months, we collected slightly more than 7,000 SMSes that we carefully annotated. This database is now ready to be used by the community. The purpose of this article is to relate the efforts put into designing this database and provide some data analysis of the main linguistic phenomena that we have annotated. We also report on a sociolinguistic survey we conducted within the project.

SMS communication as plurilingual communication

Hybrid language use as a challenge for classical code-switching categories

Etienne Morel*, Claudia Bucher**, Simona Pekarek Doehler*,
and Beat Siebenhaar**

*University of Neuchâtel / **University of Leipzig

1. Introduction: Code-switching and SMS communication

Since the first message was sent in 1990, we have witnessed an explosion of Short Message Service (SMS) use. In 2011, approximately 7.8 trillion SMS messages were sent around the globe. While younger people (adolescents and young adults) appear to be the most avid SMS users, this new way of communicating and socializing has been adopted by people of all ages working in all kinds of professions and with all levels of education. The use of more than one language in SMS communication is widespread, yet has remained scarcely explored in existing research. Plurilingual SMS practices crosscut what has classically been defined as unilingual vs. bi/plurilingual communities and unilingual vs. bi/plurilingual speakers, and are apparent in a large variety of ways. In this paper we ask: What methodological and conceptual tools are needed in order to empirically investigate code-switching in large databases of SMS communication? To what extent is the conceptual and the methodological apparatus of classical code-switching studies applicable to the investigation of SMS communication? What new issues arise from the study of SMS communication in relation to the definition of code and code-switching?

In what follows, we first present an outline of existing studies on code-switching in face-to-face and in SMS communication (§ 1). We then expose the data used for this study (§ 2). In the main body of this paper (§ 3) we discuss specific problems related to the definition of ‘codes’ in SMS communication, and ensuing issues of identifying and categorizing code-switching. We argue that the analysis of authentic SMSes calls for a broadening of current understandings of code-switching, so as to comprise not only switching between natural languages and

language varieties but also style shifts as well as switching between language and other semiotic systems (ideographic switching). We conclude (§ 4) by suggesting that the hybridization of language use that materialized in the code-switching practices of SMS users may be due to some of these users displaying their membership in a community of practice in a way that subtly navigates between localized and globalized language use and communicative values.

1.1 Code-switching and bilingual communication

For several decades now, the alternative use of two or more codes within the same communicative episode has been the object of intense debates as regards both its grammatical (what are the syntactic constraints on language alternation?) and its sociolinguistic and interactional properties (what are the social functions of alternation?). In this paper, we use the term ‘code-switching’ (henceforth: CS) as a hyperonym encompassing different types of code alternation, and we focus on the formal properties of code-switching.

Following S. Poplack (1980), the literature commonly distinguishes between inter-sentential and intra-sentential CS (sometimes, code-mixing is used for the latter, cf. P. Muysken 2000). Furthermore, at the level of intra-sentential CS, the insertion of material (following P. Auer 1996 and P. Muysken 2000 — ‘transfer’ in P. Auer 1984; ‘embedding’ in C. Myers-Scotton 1992) stemming from a language *x* within an utterance (or sentence) whose base language is *y* is distinguished from the alternation between two distinct base languages. Alternation hence corresponds to a juxtaposition, without embedding, of two different base languages. Borrowing consists in the adaptation of (mainly) lexical material to the morphological, syntactic and usually phonological structure of the host language. Borrowing can be nonce (spontaneous) or sedimented.

On a functional level, P. Auer (1984:7) defines CS as the “locally functional usage of two languages in an interactional episode”. Thus, switching carries meaning; it represents a contextualization cue (J. Gumperz 1982), and it signals the participants’ mutual orientation and indexes the social and contextual dimensions of their encounter (cf. P. Auer 1984; Li Wei 2002). From the existing literature it becomes clear that there is a complex interrelation between the forms and functions of CS. For instance, uses of CS to organise discourse appear to be typically related to CS beyond the sentence or utterance level (inter-sentential CS): such CS has been shown to be accomplished by participants in social interaction in order to index the culminating point of conversational episodes (M. Sebba and A. Wootton 1998), to mark a change in topic or in participation structure (P. Auer 1984), or to delimit a stretch of talk as indirect speech or as a side-sequence (P. Auer 1984; G. Lüdi and B. Py 2003). Intra-sentential CS appears to fulfil, at least in part, other

communicative purposes than inter-sentential CS. Insertions, for instance, may be related to enhancing the referential potential of an expression or indexing the belonging of a referential element to a specific domain of experience or a specific socio-cultural setting (cf. G. Lüdi and B. Py³ 2003; P. Auer 1984). S. Poplack (1980) suggests that intra-sentential CS implies higher language skills and is more common by 'true' bilinguals, while inter-sentential CS is more used by speakers with less bilingual skills. The findings emanating from the quoted research indicate that close scrutiny of the formal aspects of CS may provide important insights into the communicative functions of CS and enhance our understanding of the plurilingual characteristics of a given community of practice.

While the forms and the functions of CS have attracted much attention in research, existing studies have been mostly concerned with the analysis of spoken, typically interactional, data. Written plurilingual practices have remained under-researched for a long time. Only with the advent of the Internet has researchers' interest in written CS somewhat increased, and concepts stemming from the study on CS in spoken interaction have started to be applied to the study of the written genre, and in particular to CMC (cf. M. Sebba 2012: 4ff.). This application, however, raises a set of analytic problems which we address in this paper.

1.2 SMS communication as plurilingual communication?

While the social functions of SMS and aspects such as orthography, graphic icons, and the use of abbreviations, have attracted much attention in research (cf. C. Thurlow and M. Poff 2013: 164ff., for an overview), studies investigating the co-presence of two or more languages in SMS communication are still rare. The few existing studies focus on SMS practices of bilingual speakers, within bilingual/plurilingual communicative contexts, most often exploring the co-presence of English along with other languages (e.g. M. L. Bautista 2004; M. Carrier and S. Benitez 2010; I. Chiluba 2008; A. Deumert and O. Masinyana 2008; M. Haggan 2007). Particular attention has been focussed on the frequency of the use of two (or more) languages, and their effect on the length of the messages (M. L. Bautista 2004; M. Carrier and S. Benitez 2010; A. Deumert and O. Masinyana 2008).

At the current state of research, no overall explanation of CS practices has been offered by the existing literature; rather cultural differences have been evidenced in studies on CS in Kuwait (M. Haggan 2007), Nigeria (I. Chiluba 2008), Switzerland (S. Spycker 2004; S. Pekarek Doehler 2011) and Belgium (L.-A. Cougnon 2011). While these studies document a great diversity in the bilingual practices in SMS communication, mostly within bilingual communities, they say little about the formal aspects of CS, *i.e.* about the type of language units that speakers change (morphemes, words, idiomatic expressions, sentences, or stretches of more than

one sentence). Given the complex interrelation between such formal aspects of CS and issues of communicative function and bilingual community (see above), a close examination of the formal features of CS in SMS messages may centrally enrich our understanding of SMS communication across many languages, settings and communities.

2. The data and its linguistic context

2.1 Data

The Sinergia project “SMS communication in Switzerland: Facets of linguistic variation in a multilingual country”¹ is designed to explore both the language-specific grammatical properties of SMSes and patterns of CS. The project is based on the newly established SMS reference corpus for Switzerland, comprising some 26,000 original text messages from all over the country. The project sms4science.ch was launched in 2009, as a subproject of the international project [sms4science](http://sms4science.org) (cf. www.sms4science.org). From September 2009 until July 2011, 10,718 Swiss German, 7,224 Standard German, 4,628 French, 1,474 Italian, and 1,085 Romansh messages were collected. About half of the message ‘donors’ (1,316 out of 2,784) filled in a demographic online questionnaire, and therefore 20,413 messages can be related to the participant’s age, sex, education, profession, first language, SMS use, reading and writing habits, etc.

The authors of this paper are engaged in a subproject which explores the plurilingual nature of SMS communication within messages composed in French and in (Swiss-) German as a base language. The subproject is designed to contribute to a better understanding of the plurilingual nature and practice of SMS communication. It also aspires to complete our current understanding of the forms and functions of CS and to offer empirical insights into how multilingualism is concretely enacted within everyday practices in Switzerland.

2.2 The Swiss linguistic landscape

Switzerland is considered a multilingual country. This is true as a description of its linguistic landscape at the federal level (the four “national languages” are German, French, Italian, and Romansh; additionally, 9% of the population have a first language other than a national language, G. Lüdi and I. Werlen 2005: 7). However, according to the principle of territoriality, in principle almost every district (canton)

1. The project is financed by the Swiss National Science Foundation (2011–2014).

has its own official language. The situation is further complicated by the fact that in the German-speaking (and the Romansh-speaking) part of Switzerland, regional varieties rather than the official/national 'standard' languages are used in everyday life (diglossia; cf. F. Rash 1998; G. Lüdi and I. Werlen 2005). The distinction between Swiss standard German and the Swiss German dialects is manifest at all linguistic levels, from articulatory differences and different phonological systems to unequal morphology, differing syntactic possibilities, and lexica well developed in both varieties (I. Hove 2008). Speakers are aware of these differences and they have a clear ideology in separating these varieties (R. Berthele 2010). In the French-speaking part of Switzerland, by contrast, there is no diglossic situation. The traditional dialects have almost completely disappeared; with very few exceptions, the French-speaking population speaks standard French. Like in Belgium or Canada, there are only phonetic, lexical and phraseological particularities in the regional variety (A. Thibault and P. Knecht 2004: 10). One tricky issue that arises in the study of CS in SMS communication (and more generally CMC) regards code separation: what is treated as distinct codes; and to what extent can codes be unambiguously differentiated? This is what we turn to in the next section.

3. Fuzzy borders: Identifying codes and code-switches in SMS communication

One of the major challenges we face when dealing with code-switching (CS) lies in the controversial definition of what a linguistic code is. Challenges as to what a 'code' is may arise when working on oral corpora, for instance due to close phonetic, lexical and/or morphosyntactic relatedness between languages (cf. C. Alvarez-Cáccamo 1998, who investigates data comprising both Castilian and Galician). Challenges as to what a 'code' is may also arise regarding written media, for instance in the case of homography (cf. B. Siebenhaar 2006 for a distinction between Swiss German and standard German in chat data) or when we are confronted with encoding strategies such as graphic play.

The widespread language-as-code bias in CS research is typically guided by covert normative or monolingual language ideology. 'Languages', however, often do not present themselves as clearly distinguishable entities, and switches from one code to the other may be understood in terms that go beyond classical conceptions of what a language is. An eloquent example for the fuzzy borders between languages is provided by the fact that many individual words cannot be unambiguously attributed to one sole language (cf. P. Auer 2007). For instance, in written data, *chaos* or *version* can be read as English, German or French. While traditional CS research has mainly looked at instances of change from one natural language

into another, other variation phenomena (specific vocabulary use, speech styles, graphical cues) may have effects comparable to CS (M. Meeuwis and J. Blommaert 1998:80). Once we approach CS from its functional end it becomes obvious that switching into another language is just one resource amongst many for accomplishing a range of communicative purposes. For instance, within SMS communication, displaying intimacy, expressing solidarity and group identity, or simply emphasizing a phatic element is often done through language alternation (L.-A. Cougnon 2011; S. Pekarek Doehler 2011); most frequently, however, other salient language phenomena, such as the use of capitals or emoticons, serve exactly the same communicative purposes. At the same time, not every use of a language other than the base language can be interpreted as serving such purposes. In the context of SMS communication, smileys, non-standard spelling, neology and semantic shifts (cf. R. Panckhurst 2009) may have a greater impact on interaction and the negotiation of identities than the routinized use of ‘foreign’ expressions such like *sorry*, *hi* or *please*.

While recognizing the methodological advantage of focusing on switching between languages in the classical sense (French, English, Italian, German/Swiss German etc.), we stress that research on CS, and in particular on CS in CMC, needs to stay attentive to the limits of defining the relevance of a code from a researcher’s perspective. What researchers consider to be a code distinct from another may not necessarily be so for the users and vice versa. A significant example is provided by the highly routinized character of some English words or expressions within SMS communication in other languages than English (cf. S. Pekarek Doehler 2011). A quite frequent English insertion like *now* in the Swiss French corpus for instance (8 occurrences, 7 different users; see ex. 2), might be seen to be more integrated in the language than the hapax shown in excerpt (1):

- (1) *Tu fais koi mÆtnÅ?* (SMS ID 11)
 tu fais quoi maintenant?
 /tyfekwamẽtnã/
 (What are you doing now [f]?)²

2. The original excerpts are followed by an English translation, in simple parenthesis. Within the translation, the abbreviations in square brackets indicate the language of the preceding stretch of the message, be it a single item, a sentence, a part of it or more: [f] stands for French, [g] for German, [gsw] for Swiss German, [it] for Italian, [sp] for Spanish. Whenever a language is not clearly identifiable, [?] is used; in case of congruence between German and Swiss German, [g/gsw] is used; a blend of two languages is noted in the same way, for example [gsw/f] for Swiss German and French.

Where necessary for the analysis, a word-by-word gloss is inserted between the original and the translation; within the gloss, INF indicates an infinitive form; PP indicates a past participle and ACC indicates an accusative.

- (2) *Coucou lili Ça va? Alors vs etes déjà à la fete? On descend now! Gro bisous*
(SMS ID 14906)

(Hi Lili how are you? So are you already at the party? We're coming down [f]
now [e]! Big kisses [f])

Excerpt (1) contains a creative spelling of the standard French *maintenant* 'now' replacing the two French nasal vowels with the letters Æ and Å used in Scandinavian languages. The playful effect of this creativity is comparable to what is being done through CS in excerpt (2) and might in fact even top its saliency given the relative frequency of *now* in the data, which suggests a certain degree of sedimentation of the (originally) English expression within SMS communication in French.

In what follows, and based on our corpus, we consider a number of issues that represent an important challenge for the identification and categorization of the codes at play in SMS communication. Special attention will be given to problems of language attribution. We will broach these issues under the following five headings: (1) style shifting, neologisms and semantic shift, (2) pseudo-borrowings (C. Frehner 2008) and allogenisms (E. Winter-Froemel 2011), (3) homographs/homophones, (4) mixed spellings, (5) ideographic switching.

3.1 The codes at play

3.1.1 *Style shifting, neologisms and semantic shift*

Researchers often distinguish bilingual or multilingual switching between languages (code-switching or language shifting) from monolingual style shifting between different varieties of the same language (B. Bullock and A. Toribio 2009: 2). C. Alvarez-Cáccamo (1998: 2) calls for an extension of the notion of CS to monolingual speech, suggesting that language attribution alone is not necessarily relevant for the identification of CS (cf. also P. Auer 2007). Style shifting can thus be considered as a subtype of CS (P. Muysken 2000). The analysis carried out on our data so far suggests that internal language creativity as it materialises in neologisms and semantic shifts tends to fulfil comparable functions to those observed for CS between languages, such as enhancing expressivity and intimacy. Excerpts (3) and (4) show the use of code-switching between two languages while (5) and (6) illustrate style switches. The switches in all the quoted excerpts appear to fulfil highly similar (if not identical) communicative purposes.

Additionally, and for the sake of clarity, excerpt (1) contains a rewriting of the original message into standard French spelling (second line of the excerpt) plus a phonetic transcript of the pronunciation 'imitated' by the original (third line of excerpt).

Names of people have been anonymised.

- (3) *Moi aussi!:-) Ça aurait été triste de manquer une occasion de festoyer ensemble! Surtout si c'est pour toi! Baci* (SMS ID 21130)
 (Me too!:-) It would have been sad to miss an opportunity to party together!
 Especially if it is for you [f]! Kisses [it])
- (4) *Pas de soucis kiss* (SMS ID 10591)
 (Don't worry [f] kiss [e])
- (5) *Hey puce! sa va? Oui...je penre que ce merveilleux test est en salle normale... on va se marrer! Vivement ce soir... bisouilles* (SMS ID19375)
 (Hey sweetheart! How are you? Yes...I think that this wonderful test is in the normal room... that will be fun! Can't wait for tonight...kisses [f])
- (6) *Est-ce que je pourrais venir dormir ce soir?Je dois aller à Lausanne demain... :-) bisounours* (SMS ID 21868)
 (Can I sleep at yours tonight? I have to go to Lausanne tomorrow... :-) kiss³ [f])

The switches illustrated in excerpts (3) to (6) show a range of parallelisms: their sequential positioning within the message (end of message), their self-standing character as a communicative act (a goodbye formula); their function of enhancing expressivity within affective leave-taking. The intensification of intimacy beyond the mere phatic dimension is a site where other languages regularly come into play within SMSes (S. Pekarek Doehler 2011). This is shown for the Italian *Baci* in (3) or the English *kiss* in (4). Excerpt (5) however presents an example of spontaneous language internal word formation, in this case by suffixation of the noun *bisou* through *-ouille*. This suffix generally gives a familiar, popular or diminutive connotation to the noun it is attached to (as in *merdouille* or *scribouille*) and as hypocorism may give the kiss a more intimate and more personal scope. In excerpt (6) the innovation consists of a semantic shift in the use of a brand name that was originally created as portmanteau word by agglutination of the nouns *bisou* (kiss) and *ours* (bear) (*bisounours* are teddies called Care Bears in English and were mainly marketed in the 1980s). In excerpt (6) the proper noun has undergone a semantic shift and is used as a personalized affective farewell formula (cf. L.-A. Coughon 2011: 196). As far as can be seen from the analysed data, the closing sequences in SMSes thus seem to represent a slot favouring more or less personalised innovations amongst which CS is one possible option.

Similar analogies between the use of code-switches and style switches emerge in the opening sections of SMSes, as shown in excerpts (7) to (9) and (10) to (12) respectively.

3. or Care Bears.

- (7) *Schlaf gut mon amour ;)* ich ruf mal an, wenn ich vom schaffen zuhaus bin
:) kuss (SMS ID 13440)
(Sleep well [g] my love [f] ;) I'll call you when I'm back home from work:)
kiss [g])
- (8) *Ola mi amor, bisch echt widr go pfluuse. Ich lieb dich soo fescht min schazz*
[...] (SMS ID 1001)⁴
(Hi my love [sp] did you really go sleeping again. I love you so much my
dear [...] [gsw])
- (9) *Salut mon cookie chéri! [...]* (SMS ID 8509)
(Hi my [f] cookie [e] darling [f]! [...])
- (10) *Guete morge mis schnäggehüslì Sry, dass ih nid abgno han. [...]* (SMS ID 99)⁵
good morning my little-snail-house Sry, that I not picked up have.
(Good morning my little snail shell [gsw] sorry I didn't pick up the phone
[gsw].)
- (11) *Bonne nuit mon sapin. Fais de beaux rêves. Je t'aime et à bientôt:-**
(SMS ID 10521)
(Goodnight my fir tree. Have nice dreams. I love you and see you soon [f] :-*)
- (12) *Salut mon opossum!! Commen ca va? [...]* (SMS ID 22754)
(hi my opossum!! How are you [f]? [...])

Again, the highlighted elements in excerpts (7) to (12) show strong parallelisms with regard to their sequential positioning (within the opening formula of the messages), their linguistic communicative properties (terms of address), and possibly their communicative function (increasing the expressivity of affective terms of address). Excerpts (7) and (8) show typical examples for the use of romance languages in the Swiss SMS corpus within messages whose base language is (Swiss-)German;⁶ note that in (8) the entire segment *Ola mi amor* is a code-switch, and not just the term of address *mi amor*. The English *cookie* in excerpt (9) is a hapax in this use as an appellative. By contrast, excerpts (10) to (12) do not contain alternation between languages, but show instances of inventive appellatives such

4. Standard German: *Ola mi amor* [sp], bist du echt wieder schlafen gegangen. Ich liebe dich so fest mein Schatz [gsw] [...]

5. Standard German: *Guten Morgen mein Schneckenhäuschen. Sry, dass ich nicht abgenommen habe* [gsw]. [...]

6. 4 occurrences of French *mon amour* used by 4 different users and 6 occurrences of Spanish *mi amor* used by 6 different users in the Swiss German corpus; for the French subcorpus cf. S. Pekarek Doehler 2011.

as *mis schnäggehüsl* (my little snail shell) in Swiss German, *mon sapin* (my fir tree) and *mon opossum* (my opossum) in French; the analogy to the CS examples in terms of saliency but also structural embedding (greeting+my+apostrophe) and sequential positioning points at closely related functionalities.

The above observations pertain to a large number of occurrences in the SMS database for which the quoted excerpts provide only a limited set of illustrations. These occurrences provide robust empirical arguments in favour of an approach to code-switching in SMS communication that treats switches between languages and switches based on style shifts or neologisms as related rather than categorically distinct phenomena.

3.1.2 *Pseudo-borrowings and allogenisms*

The second challenge when it comes to analysing CS in SMS data relates to what are commonly called pseudo-borrowings. The SMS writer may use words that look as if they were borrowed from another language but actually do not exist in that language. C. Frehner speaks of pseudo-anglicisms (2008: 257) for words such as *handy* (mobile phone) and *checken* (to understand) that are strongly entrenched in German. The notion of pseudo-borrowing, which covers a range of heterogeneous phenomena (for a critical discussion see E. Winter-Froemel (2011: 59ff.)), is usually broached from a synchronous contrastive perspective. There may however be borrowings that seem to be pseudo-borrowings from a synchronous point of view but, at a certain point in the past, used to be 'true' borrowings. The progressive shift from 'true' to seemingly pseudo-borrowing may be due to semantic, morphological or phonetic change (e.g. the semantic shift of Spanish *sombrero* (hat) to French *sombrero* (wide-brimmed hat), compound ellipses such as French *dancing* or German *Dancing* originally related to English *dancing hall*, or the truncation in French basket from English basketball). E. Winter-Froemel (2011: 63) considers these as true borrowings. Allogenisms on the other hand are innovations within the target language that are based on procedures and units that are not part of this very language; this can be done, for instance, through attaching foreign suffixes to a target language root. Examples of allogenisms are provided in excerpts (13–15).

- (13) *Alles paletti hier. Wetter (noch) gut. Waren heute am Meer, eßen jetzt Pizza und trinken Wein. Baci Melanie* (SMS ID400)
(Everything [g] alright [?] here. Weather (still) good. Were at the sea today, now eating pizza and drinking wine [g]. Kisses [it] Melanie)

- (14) *Saletti spaghetti wie gahts? Wötsch hüt verbicho? Mir wäred am halbi 4i dihei!*
(SMS ID15924)⁷

(Hi there [?] how are you? Do you want to come by today? We'll be at home at half past 3 [gsw]!)

- (15) *No **problemo**, je m'en occupe! Bisous, valentine* (SMS ID9368)
(No problem [?], I'll take care of it! Kisses, valentine [f])

Excerpts (13) and (14) show allogenisms created by means of the suffixation of the Italian suffix *-etto/i*. Although allogenisms cannot be considered as CS in the same sense as switches from one language to another, they may have related functions as code-switched elements, and in particular as insertions. In (13) the sender is most likely spending her holiday in Italy (see the deictic *hier*, 'here') and uses an allogenism showing incipient sedimentation in the language (it has found its way into dictionaries)⁸ for indexing her current location. The very same functional use has been documented by S. Pekarek Doehler (2011) for more classical forms of CS in the French SMS subcorpus. Excerpt (14) contains a less sedimented allogenism which is not included in dictionaries, yet recurrently found in the Swiss German subcorpus.⁹ Here again the use of the allogenism has a strong parallel to what S. Pekarek Doehler (2011) found for CS in the French SMS subcorpus; these are often related to formulaic greetings such as *ciao bella* or *hey buddy*. As to excerpt (15), it shows the imitation of a romance language by o-suffixation of an international word, which however would run *problema* in Italian or Spanish (*problème* in French). *No **problemo*** is often cited as instance of mock-Spanish or junk-Spanish in a Northern American context, as "appropriation of Spanish in white spaces" (A. Schwartz 2011). In our sociolinguistic context it is likely to be either an imitation of Italian or Spanish or a borrowing via American English, as it is highly established there.¹⁰ The imitation of Italian¹¹ and other migration languages (such as Portuguese in the French-speaking part of Switzerland) raises interesting questions as to identity construction and group affiliation that have been at the core of CS studies on spoken interaction. While a systematic analysis of these phenomena

7. Standard German: Saletti spaghetti [?] wie geht's? Möchtest du heute vorbeikommen? Wir wären um halb vier zu Hause [gsw]!

8. http://wortschatz.uni-leipzig.de/cgi-portal/de/wort_www?site=208&Wort_id=17206610. We found 10 uses by 10 different users in the Swiss German SMS subcorpus.

9. 7 occurrences produced by 4 different users in our corpus.

10. It became popular in the late 80s by the A. Schwarzenegger Terminator II movie.

11. See also the imitation of an Italian migrant's Swiss German by Viktor Giacobbo, a popular late night show presenter: <http://www.youtube.com/watch?v=HdQOaHr9Ltg>.

in SMS communication remains to be undertaken, the quoted excerpts again suggest that the borderline between allogenisms (and possibly pseudo-borrowings) on the one hand, and what is considered as classical instances of CS may not be clear-cut as it seems at first sight, and raises the critical issue as to what the codes (and the distinctions among these) are for the SMS users themselves.

3.1.3 Homographs/homophones

The third challenge we face when trying to identify the code-switches and languages present in a text message are homographs. Some words are spelled alike in different languages (*chaos*) and even typical SMS abbreviations like *n8* can be meant to transcribe German ‘n-acht’ (*Nacht*), French ‘n-huit’ (*nuit*) or English ‘n-eight’ (*night*). In excerpt (16) punctuation may possibly indicate the likely switch point, and therefore an attribution to French seems most likely (*n8* standing for ‘nuit’). Yet, in combination with the German apostrophe *spatzilein*, a switch between *bonne* and *n8* is also conceivable, in which case *n8* would stand for ‘Nacht’.

- (16) *bonne n8, spatzilein!!* :* :* (SMS ID33)
(good [f] n8 [?] little sparrow [g] !! :* :*)

Homophonous diamorphs in oral code-switching (P. Auer 2007; M. Clyne 1967; P. Muysken 2000:133f.) represent a key problem for language tagging, and the absence of pronunciation cues makes this a particularly delicate issue for the study of written data. However, as P. Auer (2006) correctly notes, speakers who code-switch can actually be interested in blurring the boundaries between languages. This may aim at smoothing the language transition but possibly also at playing on the hybridization of languages and identities in order to display membership in a globalized community of practice. Such hybrid language use is further illustrated in excerpt (17):

- (17) *Ça va, le w.e. was worth l'effort* (SMS ID9979)
It goes, the [f] week-end [g/f/e] was worth [e] the [f] effort [e/f/gsw]
que je fais maintenant... :) [...]
that I make now [f]...:) [...]
(I'm ok, the weekend was worth the effort that I'm making now...) [...]

Excerpt (17) shows two homographs: At the beginning of the English stretch the abbreviated word *week-end*, a highly frequent loan word in French, potentially triggers the switch. At the end of the stretch, the ambivalence of the English-French homograph *effort*, combined with the determiner *le* (here abbreviated to *l'*) is a turning point for the switch back to French. The hybrid status of the words in question is specific to the written medium (pronunciation would typically disambiguate language attribution issues). However, as mentioned earlier, what might

appear as ambivalence to the researcher may serve as an important resource for the language user. The hybrid status of the elements quoted above may be exploited by SMS writers for the purpose of blurring the boundaries between languages as part of a trans-linguistic code that is symptomatic of an internationalized and globalized communicative culture, and may possibly be used for indexing the writer's membership to the related community of practice. While the pragmatic functions of such occurrences as well as their possible status as CS call for a detailed case-by-case analysis, when it comes to analysing large databases they are most usefully dealt with in an initial step (*i.e.* in CS/language tagging) as what they are: namely linguistic forms that can be attributed to more than one language.

A special problem concerning homographs arises with the distinction between standard German and Swiss German dialects. From the linguist's perspective these are two varieties of one language. From the speaker's point of view, the two varieties are clearly separated, to a point that many Swiss German speakers state standard German as their first foreign language at least for spoken language (R. Berthele 2004; H. Christen *et al.* 2010; I. Hove 2008; G. Lüdi and I. Werlen 2005). Therefore, a switch between the standard and the dialectal variety is often functional and can be seen as a CS. Due to their close relatedness, the attribution of a word or expression to either variety is often impossible. In excerpts (18) to (20) words in bold can unambiguously be attributed to a Swiss German dialect, underlined words are unambiguously standard German, and the rest can either be dialect or standard.

- (18) *Hey **läuft** no was **beder**? Mer send **erscht** ez fertig worde... [...](SMS ID235)¹²*
 Hey goes still something at yours? We are only now finished...
 (Hi, is there something going on at yours? We have just finished now...)
- (19) ***Hoi** joshua Komme sehr knapp unser bus steht still. Bis **när** (SMS ID31)¹³*
 Hi Joshua. Come very short our bus stands still. Until after
 (Hi Joshua. I will arrive just in time. Our bus is standing still. See you.)
- (20) *Weiss **no** nüd, ob ich am Mi in Luzern bin — meld mich noch. Glgl*
 (SMS ID244)¹⁴
 Know yet nothing, whether I on Wed in Lucerne am — tell still. Greetings
 (I don't know yet, if I will be in Lucerne on Wednesday — I'll get in touch.
 Greetings.)

12. Standard German: Hey [g/gsw] läuft noch [gsw] was [g] bei dir? Wir sind erst jetzt [gsw] fertig [g/gsw] geworden... [gsw] [...]

13. Standard German: Hallo [gsw] Joshua Komme sehr knapp unser Bus steht still. Bis [g/gsw] nachher [gsw]

14. Standard German: Weiß [g/gsw] noch nicht [gsw], ob ich am Mi in Luzern bin — melde mich [g/gsw] noch [g]. Glgl [g/gsw]

Excerpt (18) can be reasonably interpreted as a message written in Swiss German: any of the items of the messages can be read as Swiss German (though some of them can be read as either Swiss German or standard German). In excerpt (19) there are Swiss German and standard German words and some that can be either Swiss German or standard German. The dialectal items are at the edges of the message in greeting formulas, this allows for an interpretation as CS due to their distinct function. By contrast, excerpt (20) is particularly tricky. Most words can be either standard or dialect, but three words with no apparently specific communicative function have conflicting spelling attributes, so the message as a whole cannot be seen as dialectal nor as standard German. As these three words do not seem to have a specific function within the message there is no justifiable possibility to interpret the change of the variety as a CS. Moreover, if the message were split up into a dialectal and a standard German part, the location of the switch could not be clearly substantiated. In this specific case the ambiguous use of the two varieties could be seen as a step towards what P. Muysken (2000) calls congruent lexicalization, where the different codes are characterized by structural equivalence and an abundance of homophonous words which make back-and-forth switches possible.

3.1.4 *Mixed spellings and morphological variants*

Non-standard spelling and graphic variation is probably one of the characteristics of text messaging that has received the greatest deal of attention in recent research (for French: J. Anis 2001; 2007; M.-J. Béguelin 2012; C. Fairon *et al.* 2006: 31–47; R. Panckhurst 2009). M.-J. Béguelin (2012), analysing the sms4science.ch French sub-corpus, identified a set of different spelling uses, such as partial or complete scriptio continua (omission of spaces), generalization of certain graphemes for economic (and other) reasons (such as *k* for *qu*), syllabograms through phoneticisation of numbers and letters (as in the above-mentioned *n8*), consonant skeletons (e.g. *dsl* for *désolé*, sorry) and other types of abbreviation (e.g. truncation). For research on written CS, such neography is particularly significant in the case of novel spelling habits which can be ascribed to the influence of foreign alphabetic conventions. These switches between graphematic systems can be pragmatically motivated. V. Hinnenkamp (2008), who looked at online language behaviour of adolescents with ethnic Turkish background in German chat rooms, shows that mixing languages through orthographic play, such as transcribing *Deutsch* as *Doyc* or *Doitsch*, can be intentional and may bear pragmatic meaning (cf. also A. Jaffe *et al.* 2012).

Mixing spelling conventions can go in two distinct directions: the base language can be transcribed with the conventions of another language or the inserted language can be adapted to the spelling of the base language. In principle, a third language could come into play, though we did not find any such examples in the data (think of a possible occurrence of **batschi* in a French SMS, where an Italian

word is transcribed through a German spelling convention, as shown in the 'sch'). Excerpts (21) to (23) show instances of mixed spelling:

- (21) *Bijour m'sieur! Alors comment ajd?! L'estomac est à l'endroit ou à l'envers?! De sortie chez Mimi? Gros bizooX* (SMS ID15549)
(Hi Sir! So how are you today?! Is your stomach the right way up or upside down?! Are you going out to Mimi's? Big kisses [f])
- (22) *heii shatz ja wemer öpis mache zeme? wen trefemer eus? lys.<3* (SMS ID2512)¹⁵
(Hi darling well do you want to do something together? When shall we meet [gsw]? lys [e]. <3)
- (23) *Hey Hesch morn frei?Wenn ja ech au:-) chönte ja ad mäß gah. Odr bisch scho gsi? Lysm kuzzy* (SMS ID2080)¹⁶
(Hey [e] are you off work tomorrow? If so me too:-) We could go to the fair. Or have you already been there [gsw]? Love you so much [e] kiss [gsw])

Excerpt (21) shows an example of neography related to the influence of a language other than the message's base language. Here, the French word *bisou* owes its transcription to a typically English spelling of the closed back vowel /u/, namely 'oo'. The word ends on an ideographic X which most often represents a kiss. Interestingly though, -x is also the plural suffix for some French nouns ending on -ou (this is however not the case for *bisou*). This linguistic play suggests an increased metalinguistic awareness on the part of the writer (C. Tagg 2009), who is possibly engaged in what D. Crystal (2008:73) calls putting up the ante, *i.e.* pushing innovations further in the sense of poetic *æmulatio* in order to display a specific type of competence as an SMS writer.

Excerpt (22) illustrates a feature recurrently found in the data, namely the influence of English on the transcription of the post-alveolar fricative /ʃ/ in German and Swiss German text messages: the German spelling *sch* is often replaced by *sh* (cf. C. Frehner 2008:250). This may of course be done for the purpose of saving space, yet explanations in terms of language economy have proven to fall short of accounting for variation phenomena in SMS communication (see § 1.2. above). In excerpt (23) *kuzzy* represents an English stylization of the German neologism *Kussi* (from *Kuss*, kiss), involving the replacement of *s* by *z*; the replacement of the final -i by -y, though without phonetic consequences, gives the word an English graphematic touch, as -y is a frequent adjectival suffix in English (like in *fuzzy* or

15. Standard German: Heii Schatz, ja wollen wir etwas zusammen machen? Wann treffen wir uns? [gsw] lys [e]. <3

16. Standard German: Hey [e] hast du morgen frei? Wenn ja ich auch :) Wir könnten ja auf die Messe gehen. Oder bist du schon gewesen [gsw]? Lysm [e] kuzzy [gsw]

jazzy) and as German words ending on -y are typically English borrowings (*hobby*, *baby*, *cowboy*) or allogenisms (*handy*).

It is also important to note what M.-J. Béguelin (2012) has pointed out as one of the most striking phenomena of SMS communication, *i.e.* the transcription of morphological variants which usually would be neutralized through writing and which results in writing that does not follow the rules of spelling but imitates a certain pronunciation. Excerpts (24) to (26) present examples:

- (24) *Yo f., wi géts?romscht du a lentrainement?si ta lvoiture, puis taccompagner?*
(SMS ID 16583)¹⁷
Yo f., how goes-it? come you to training? if you.have a car, can you.ACC-join?
(Yo f. [e] how are you? Are you coming [gsw/f] to training? If you have a car,
can I join you [f]?)
- (25) *Coucou.Tu rentre à loz ce week end? À plouche...* (SMS ID14910)
Hi. You come.back to Laus this weekend? Until later.
(Hi. Are you coming back to Lausanne this weekend [f]? See you later [f/?])
- (26) *Boshur! Hei du, au wen du odr/und i naniq qa/qasH qo penne: sHlaf quet! Bis morn uf dä wanderiq! Hdumfl :-** (SMS ID6142)¹⁸
(Hello [f/gsw]! Hey you, even if you or/and I cannot fall asleep: sleep well!
See you tomorrow at the hike! Love you so much [gsw] (xxx) :-*)

The transcription of morphological variants can result in playful texts where the writer seems to imitate a certain pronunciation, for example of Swiss German as in excerpt (24); especially noteworthy is the use of r to transcribe the emblematic voiceless uvular fricative /χ/ in *chunsch du* (/χʊnʃdu:/ ‘are you coming’). In excerpt (25) foreign pronunciation of French is imitated: *À plouche* stands for *à plus* (/ap-lys/ ‘see you later’) and is reminiscent of the emblematic pronunciation of French by certain Portuguese-speakers (close front rounded vowel /y/ pronounced as closed back vowel /u/; alveolar fricative /s/ as post-alveolar fricative /ʃ/). Excerpt (26), finally, shows the imitation of a German pronunciation of French *bonjour* by annulment of nasality and replacement of voiced fricative through its voiceless variant (/bɔʃur/ instead of /bɔʒur/).

Just as the hybrid language use documented in Section 3.1.3., the use of mixed spellings is a highly significant phenomenon, particularly associated with SMS communication, and possibly indicative of a voluntary (or in some cases routinized)

17. Standard German/normalized French: Yo f. [e], wie geht es? Kommst du [gsw/f] à l’entraînement? Si tu as une voiture, puis-je t’accompagner [f]?

18. Standard German/normalized French: Bonjour [f/gsw] Hei du, auch wenn du oder/und ich noch nicht schlafen gegangen sind: Schlaf gut! Bis morgen auf der Wanderung! Hdumfl [gsw] :-*

play with the borderlines between languages, and the display of the SMS writer's membership in an internationalized and globalized community of practice.

3.1.5 Ideographic switching

A further issue in defining the codes present in SMS communication is one that bears important theoretical implications. If we admit that switching is possible between different types of graphematic conventions and even between alphabets, we are called to consider the use of ideography as possible CS. While ideography can be used to counterbalance the absence of contextual and suprasegmental information (M.-J. Béguelin 2012), sometimes the integration of emoticons into the very grammatical structure of a sentence strongly recalls the process of insertion of elements from one natural language into utterances or sentential patterns made up of other languages. This is why we suggest using the term ideographic switching to describe the phenomenon illustrated in the following excerpts:

- (27) *Coucou ma Tironia! Cmt vas-tu?Ce WE à Lisbonne? Et Alexandra? Je viens d'apprendre que nous aurons l'appart au 1er janvier je suis :-)! [...]*
(SMS ID11651)
(Hi my Tironia! How are you? Are you in Lisbon this weekend? And Alexandra? I've just heard that we will have the apartment for the first of January I am [f] :-) [...])
- (28) *Yeah baby! J'suis trop happy, tu verras comme c'est cool!!! Tèm fort aussi:) Tu me manques...*
(SMS ID 17349)
(Yeah baby [e]! I'm too [f] happy [e], you'll see how cool it is!!! Love you too very much:) I miss you [f])
- (29) *Ça yé bien arrivée! U.du??C t hüere güet, merci+++ u bonne nuité gros bisous! Bis am mäntig, dini M.*
(SMS ID1780)¹⁹
(So I arrived safely [f]! And you [g/gsw]?? It was [f] really great [gsw], thanks+++ [fr] and [g/gsw] have a good night and big kisses [f]! See you on Monday, your M.[gsw])
- (30) *oh schnuggiputz hallo,i miss you! Du hesches bald gschaftt :-) vernünftig bisch mitem zug gange, super! Cu soon :-**
(SMS ID131)²⁰
(Oh my darling hi [g/gsw], I miss you [e]! Soon you will have made it :-) you were right to go by train, super [gsw]! See you soon [e] :-*)

19. Standard German/normalised French: Ça y est bien arrivée [f]! Und du [g/gsw] ?? C'était [f] sehr gut [gsw], merci [f] und [g/gsw] bonne nuit et gros bisous [f]!

20. Standard German: Oh schnuggiputz hallo [g/gsw], i miss you [e]! Du hast es bald geschafft :-) Vernünftig bist du mit dem Zug gegangen, super [gsw]! Cu soon :-* [e]

- (31) *Guet gange bim zahni? Bin am 4i bim chiro. Wo send ihr? Zoo? Muchos besos f*
(SMS ID10150)²¹

Well went at the dentist? Am at 4 at the chiropractor. Where are you? Zoo?
Many kisses f.

(Did it go well at the dentist? I'm at the chiropractor at 4. Where are you
[gsw]? At the zoo [g/gsw]? Many kisses f [sp])

Compare for instance excerpts (27) and (28) where the verb complement of *je suis* 'I am' consists of the emoticon :-) (27) and the English adjective *happy* (28) respectively. Note also the graphic intensifier in excerpt (29), which possibly replaces the adverb *beaucoup* 'a lot'. Moreover, the frequent syntactically independent insertion of emoticons like :-* (depicting a kissing face) at the end of messages like in excerpt (30) parallels the linguistic creativity recurrently observed in closing formulas where CS is frequent (31).

All five phenomena that we have discussed in this section (neology/semantic shifts, allogenisms, homophone/homographs, mixed spellings and ideographic switching) represent major challenges for the categorization of the codes at play in CS within SMS communication. As suggested in this section, all of these phenomena are part of a wide range of frequently used resources involving the alternation and hybridization of codes, some of which may be specific to this medium of communication.

3.2 The limits of traditional CS categories

A second set of challenges that arise when it comes to analysing CS in SMS data relates to the classification of different grammatical forms of CS. The traditional formal categories of CS do not seem to account for the full array of forms of CS found in the data. As mentioned in the introduction, following S. Poplack's (1980) classic study, the literature commonly distinguishes between inter-sentential and intra-sentential code-switching (sometimes, code-mixing is used for the latter: P. Muysken 2000). Furthermore, different types of CS are distinguished at the level of intra-sentential code-switching (P. Auer 1996; P. Muysken 2000; C. Myers-Scotton 1992):

- a. insertion refers to the 'embedding' (C. Myers-Scotton 1992) of material stemming from a language *x* within an utterance whose base language is *y*. In (32), the English or French (it is a homograph) plural noun *parents* is embedded in a clause structure in Swiss German:

21. Standard German: Gut gegangen beim Zahnarzt? Bin um 4 beim Chiro. Wo seid ihr? Zoo [gsw]? Muchos besos f [sp]

- (32) *He wemmer öpis choche oder hesch scho gesse? Mini **parents** sind au weg*
(SMS ID1393)²²

(Hey do we want to cook something or have you eaten already? My [gsw]
parents [e/f] are also away [gsw])

- b. alternation refers to the switching between two distinct base languages (cf. P. Muysken 2000). Alternation hence corresponds to a juxtaposition, without embedding, of two different languages. In the complex utterance shown in (33), the first sentence presents a case of alternation, where the first clause is produced in French, whereas the that-introduced subordinate clause is in Swiss German. After that follows a simple sentence in Swiss German (*Bin wahrschindli so uf di 10ab zwölfi det*) and then a compound sentence (*Bis dänn et dsl pour le retard*) where the first independent clause is in Swiss German and the second in French.

- (33) *Je viens de rater mon train ce qui veut dire :0) daß du chli uf mich wartä
muäsch... Bin wahrschindli so uf di 10ab zwölfi det ... Bis dänn et dsl pour le
retard...* (SMS ID32)²³

(I just missed my train which means [f] :0) that you will have to wait for me
a little bit... I will be there at about 10 past twelve ... See you then [gsw] and
sorry for the delay [f] ...)

- c. borrowing — a complex phenomenon we will not be able to discuss in depth here — basically consists of the adaptation of (mainly) lexical material to the morphological and phonological structure of the host language (S. Poplack 1980:583f.; P. Muysken 2000:69). Borrowings can be more or less routinized and accessible to the speakers' community (P. Muysken 2000:71); the integration into a dictionary is the ultimate result of the sedimentation process. Further criteria for identifying borrowings may be semantic change (E. Winter-Froemel 2011:21) and orthographic integration (T. Spilioti 2009). Consider the following examples:

- (34) *Hallo! Voulez vous **luncher** avec moi hüt?* (SMS ID6417)

Hello! Want you lunch.INF with me today?

(Hi [g/gsw]! Do you want [f] to have lunch [f/e] with me [f] today [gsw] ?)

22. Standard German: He wollen wir etwas kochen oder hast du schon gegessen? Meine [gsw] parents [e/f] sind auch weg [gsw]

23. Standard German/normalized French: Je viens de rater mon train ce qui veut dire [f] :0) dass du ein wenig auf mich warten musst...bin wahrscheinlich so auf 10 nach zwölf dort... Bis dann [gsw] et désolé pour le retard [f] ...

- (35) *Pfff il est tjrs surbooké! ;)* (SMS ID13371)
 Pff it is always overbook.PP
 (Pfff it is always [f] overbooked [f/e] ! ;))

In excerpt (34) the English radical *lunch* (used in English as a noun, but here as a verb) is combined with the French inflexional morpheme *-er* (i.e. the infinitive form), and in (35) the English verb *book* is prefixed by the French *sur* (here: over-) and affixed by the inflectional morpheme *-é*, indexing the past participle.

As clear-cut as the quote excerpts might seem, in the data we find many cases where the distinction between alternation, insertion and borrowing is far from obvious.

3.2.1 *Established borrowing or code-switching?*

The first problem lies in the identification of borrowings. Morphological integration does not represent a relevant criterion for certain adverbs or for singular nouns; phonological integration cannot be observed in written data. The integration of a word into official lexicography has the merit of offering a methodologically clear-cut criterion for distinguishing between established borrowing, spontaneous borrowing, and CS. However, this criterion removes from the picture those words or expressions that may not have found their way into dictionaries yet, but seem to be undergoing a process of routinisation within a given community of practice. Consider the following excerpts:

- (36) *Coucou! Oui je suis à l'université today et g je dvd. Gros bisous.* (SMS ID13347)
 (Hi! Yes I am at the university [f] today [e] and I have the DVD. Big kisses [f].)
- (37) *Je pars now, à toute, bisous mon cœur* (SMS ID21647)
 (I'm leaving [f] now [e], see you soon, kisses my dear [f])
- (38) *[...] has au geil gefunde. Bis next week* (SMS ID14472)²⁴
 (I also thought it was great. Till [gsw] next week [e])

The *today* in excerpt (36) and the *now* in excerpt (37) are not established loan words in French, but their recurrence in the data²⁵ points to their status as loan words in statu nascendi for a certain community of practice. Other instances of such incipient sedimentation can be found in the Belgian sms4science corpus: *hi*, *sorry*, *next* or *happy* (L.-A. Cougnon 2011). Excerpt (38) presents an example for the Swiss German corpus.²⁶

24. Standard German: [...] habe es auch geil gefunden. Bis [gsw] next week [e]

25. *Today*: 12 occurrences; 10 different users; *now*: 8 occurrences; 7 different users.

26. *Next week*: 12 occurrences; 11 different users.

3.2.2 Fixed formulae

A second difficulty with applying the classical formal categories of CS to SMS data lies in the frequent occurrence of isolated, syntactically and pragmatically free-standing words or expressions (especially in greeting and closing sequences) that can neither be considered as insertions nor as alternations.

- (39) *Hi! C'est pas grave! J'ai bien. [...]* (SMS ID10136)
(Hi [e]! It's not a problem! I'm ok [f].)
- (40) *Jariv 10min, sorry* (SMS ID12715)
(I'll be there (in) 10 minutes [f], sorry [e])
- (41) *scusaa..ha sho pfüselet, du no lang gfeshtet?en guete und so* (SMS ID8814)²⁷
(Sorry [it] I was already asleep, have you been celebrating for a long time?
Enjoy your meal and everything [gsw])
- (42) *Clara bonjour -entschuldige- je serais un peu plus tard- [...]* (SMS ID23114)
(Clara hello [f] — sorry [g] — I'll be a little bit late [f] — [...])
- (43) *J t ai battu! yeah! j ai monté la pente plus vite que toi;-) [...]* (SMS ID12938)
(I beat you [f] ! yeah [e] ! I got up the slope faster than you [f] ;-) [...])

A. Backus (2003) considers such formulae as non-prototypical cases of insertional CS. However, what syntactic frame are they inserted in? They actually appear to be free-standing, independent utterances, built of linguistic material that is not embedded in any surrounding material. In the quoted excerpts their free-standing status is further evidenced by punctuation. Note that the quoted items are also pragmatically free-standing, as it were, each of them accomplishing a pragmatic act in itself (such as greeting, welcoming, excusing, etc.). While the fact that they are not embedded points to features usually attributed to alternation (P. Muysken 2000:231), the lack of morphosyntactic structuring qualifies the quoted cases as different from what is commonly known as alternations between two base languages. Given the above, and given the remarkable frequency of such formulaic segments in the data (S. Pekarek Doehler 2011), they call for the need to establish a separate formal category of CS, distinct from both alternation and insertion.

Interestingly, the above observations relate not only to isolated words but also pertain to fixed expressions as indicated by R.W. Langacker (1991). In the data, there are multimorphemic sequences that structurally may look like alternations but fundamentally differ from typical cases of alternation as quoted in (33) above, because they seem to be produced as fixed expressions, *i.e.* as grammatically

27. Standard German: *scusa* [it].. *habe schon geschlafen, hast du noch lange gefeiert? Einen Guten und so* [gsw]

unanalysed sequences. They form what has been called formulaic sequences (A. Wray 2002), also known as chunks, or what A. Backus describes as long words (2003:85).²⁸ Their main characteristic is their idiomaticity, *i.e.* their (a) non-compositionality: the fixed expression builds a semantic unit, where the overall meaning prevails over the combination of the components' respective meanings. Other typical features of formulaic sequences are (b) recurrence, (c) irregular morphosyntax, (d) pragmatic function and (e) phonological reduction (A. Backus 2003:90f.).

- (44) *heeeej.. Kathleen..what's up??en schöne tag gha??wa verzellsch so?? [...]*
(SMS ID708)²⁹

(Hey [?] .. Kathleen..what's up [e] ?? had a nice day?? Any news [gsw]? [...])

- (45) *Et non! Je dormais :-)... Je bosse a la Maladiere toute la journée! Si tu veux boire un bon café et faire tes courses, you're welcome! Bec et bonne journée*
(SMS ID9905)

(If you want to drink a good coffee and go shopping [f], you're welcome [e]!
Kiss and have a nice day [f].)

- (46) *Hehe, muen mi erschma vo hüt erhole, ha viel“dezuegelernt”; Lg & gn8 Ps: ma luege denke aber ehner weniger..We'll see!*
(SMS ID138)³⁰
(Hehe, first I have to recover from today, I learned a lot of new things;
Greetings & good night P.S.: let's see but probably not [gsw] ... We'll see [e]!)

The switched stretches in excerpts (44), (45) and (46) can easily be identified as formulas inasmuch as they correspond to most of the criteria cited above. The criterion of non-compositionality is fulfilled for *what's up* in excerpt (44) as the meaning of the whole expression cannot be deduced from the literal meaning of its components (“what is up?”); the same applies to *you're welcome* in (45), used here as an invitation (you are welcome to join me). The expressions quoted in (44), (45) and (46) are found several times in the data.³¹ All three examples provide instances of graphically transcribed phonological reductions and all three of them are marked as self-standing units through the use of punctuation.

28. A. Wray (2002:9) gives the plethora of possible terms used in research to describe aspects of formulaicity.

29. Standard German: heeeej.. Kathleen..what's up [e] ?? Einen schönen Tag gehabt??Was erzählst du so [gsw]?? [...]

30. Standard German: Hehe, muss mich erst einmal von heute erholen, habe viel “dazuegelernt” ;) Lg & gn8 Ps: mal schauen denke aber eher weniger[gsw].. We'll see [e]!

31. In the Swiss German subcorpus, for instance, there are three occurrences of *what's up* and three occurrences of *we'll see*.

Excerpts (44) to (46) represent what C. Fillmore *et al.* (1988: 505) called substantive idioms, *i.e.* formulaic expressions which are lexically completed and which contain no open slot. By contrast, (47) to (50) represent formulaic expressions that are lexically less fixed, comprising an (optional) open slot that can be filled with different lexical material (formal or lexically open idioms):

- (47) *Eh ciao amore . bish fit?he hüü ish ya dä börner xsi midm bill. [...]*
(SMS ID4844)³²

(Eh [?] hi my love [it]. are you fit? hey today it's been great with Bill [gsw]
[...])

- (48) *ciao,ja ich muen mis lebe ganz noi ufgleise,bin single [...]* (SMS ID5091)³³
(hi [it] yes I need to put my life on a new track, I'm single [gsw] [...])

- (49) *ciao bella. du, nögst versuech: ;-)* was mached ihr am SA obig, [...]
(SMS ID4323)³⁴
(hi beautiful [it], you, next try: ;-) what are you doing on Saturday evening
[gsw], [...])

What is particularly interesting from the perspective of CS is that the slot in such lexically open expressions can be filled with languages other than the rest of the expression:

- (50) *Hi my dear! Wie verlief die OP? Darf ich dich morgen abend besuchen kommen? Biiig-huge hug!* (SMS ID4449)

(Hi my dear [e]! How did your surgery go? Can I come to visit you tomorrow evening [g] ? Big huge hug [e] !)

- (51) *Hi sockeloch! Jaaa isch au schomal besser gange...bin chränk! Und han hüt schowider nurno die erst reihe verwütscht! Da chani jez umehueste ;)*
(SMS ID4768)

(Hi sock-hole! Well there were times I felt better...I'm sick! And today I only got the first row again! I can cough around here now [gsw];))³⁵

32. Standard German: Eh ciao amore [it]. Bist du fit? He heute ist ja der Burner [i.e. der Hit, der Knüller] gewesen mit dem Bill [gsw] [...]

33. Standard German: Ciao [it], ja ich muss mein Leben ganz neu aufgleisen, ich bin single [gsw] [...]

34. Standard German: Ciao bella [it]. du, nächster Versuch: ;-) was macht ihr am SA Abend, [gsw] [...]

35. Standard German: Hi [e] Sockenloch! Ja ist auch schon einmal besser gegangen...bin krank! Und habe heute schon wieder nur noch die erste Reihe erwischt! Da kann ich jetzt umherhusten [gsw] ;)

Excerpts (50) and (51) contain the formula “Hi (x)”.³⁶ In (50) the slot is filled with an English appellative whereas in (51) it is filled with a creative Swiss German apostrophe (my socks’ hole). The formulaic sequences found in the data thus appear to be located on a continuum from highly fixed and conventionalized units, through “loosely connected but collaborative elements” (N.C. Ellis 2012), to highly creative occurrences.

3.2.3 *So what?*

In what way are the observations presented in this section centrally relevant for the study of CS in SMS communication? The fact that alternation seems to be quite rare in our corpus and that CS often takes the form of an insertion of routinized words or of a self-standing formulaic expression (cf. S. Pekarek Doehler 2011), sheds light on some aspects of the community/communities of practice standing behind the data. A quantitative picture of the types of CS in a given set of data may provide clues as to which groups of writers prefer specific CS routines and which CS routines are preferred for which functions. The systematic research on a multi-language based corpus may also provide clues for differences and similarities between languages and speech communities within the globally new communication form of SMS (see § 1.1. above).

For instance, the quasi-absence of ‘true’ alternation, and the high frequency of formulaic, often highly internationalized and largely transparent formulaic expressions, suggests that the community of practice that we are analysing does not consist of bilinguals in a classical sense (cf. S. Pekarek Doehler 2011); rather, the formal types of CS and their frequency display patterns of minimal bilingualism (J. Androutsopoulos 2007). The SMS writers appear to use ‘controllable’ bits of other languages that tend to be accessible to the community, and they do so possibly “regardless of degrees of linguistic competence or ethnolinguistic affiliation” (J. Androutsopoulos 2013: 686; cf. also N. J. Jørgensen 2008).

36. In the Swiss German subcorpus the greeting formula *Hi (x)* occurs 158 times; most often the open slot is filled with a proper name (*Hi Linda*) (81 occurrences); 36 times it isn’t filled at all; 26 times it is filled with a German appellative (*Hi sockeloch*); 11 times with an English appellative (*Hi sister*) or an English adverb (*Hi there*) twice with an Italian appellative (*Hi bella*); in the French subcorpus *Hi (x)* only occurs 13 times: mostly in its simple version without filled slot (*Hi*) (10 occurrences), once with a name, once with a French (*Hi mec*) and once with an English appellative (*Hi baby*).

4. Conclusion

In this paper we set out to discuss the issues that arise for identifying and categorizing codes and code-switching in SMS communication. The empirically-based discussion presented here was designed to provide a basis for research on the communicative functions of CS in SMS communication, and to shed light on some of the as yet little-explored formal features of this mode of communication. While a few of the quoted observations pertain to the specific situation of Switzerland, or more generally dialect-standard diglossias (e.g. the tricky differentiation between standard and dialect), many of the above may be relevant for a large set of linguistic and cultural contexts.

The quoted observations call for a revision of some of the formal (and possibly functional) categories of CS known from the existing literature. In particular, they outline the need (a) to broaden our understanding of what a code (language, variety, etc.) is — and accordingly of what a CS is, and (b) draw our attention to the as yet little-explored hybridization of the very notion of language in language use.

With regard to (a), we have placed the very notion of code under critical scrutiny. We have argued on empirical grounds for a broad understanding of CS that comprises switching between natural languages and language varieties along with style shifts as well as switching between language and other semiotic systems (ideographic switching). Seeing switches between codes in relation to each other within their local ecology allows us to account for CS as part of the myriad of resources communicators put to work to accomplish their communicative purposes; and only such an understanding is able to pay tribute to the impressive creativity of language games that are emblematic of many modes of communication, including SMS communication.

Among these, and with regard to (b), hybrid forms of language use that blur the boundaries between what we commonly call languages (homographs, mixed spellings or the incipient sedimentation of items or formulae from 'other' languages) may be seen as a key feature of SMS communication. Such hybridization of language may possibly be one of the ways SMS users display their membership in a community of practice that is subtly balanced between localized and globalized language use and communicative practice.

This very point deserves close qualitative analysis, as way of addressing the sociologically central question as to how SMS communication is becoming one site where the tension between localized and globalized practices, which is a pervasive feature of today's social worlds, is played out. This analysis is the next step in our own work. The critical scrutiny of what can be considered as a code in SMS communication, and how codes are treated as related, distinct or diffuse by the SMS users themselves, provides a basis on which the functions of CS can be soundly

analysed as part of an investigation into how SMS users perform community and identity dynamics, and how, thereby, they navigate between individualization and community construction and/or affiliation.

References

- Alvarez-Cáccamo, Celso. 1998. From 'Switching Code' to 'Code-Switching': Towards a Reconceptualization of Communicative Codes. In *Code-Switching in Conversation. Language, Interaction and Identity*, P. Auer (ed.), 29–50, London/New York: Routledge.
- Androutsopoulos, Jannis. 2007. Bilingualism in the Mass Media and on the Internet. In *Bilingualism: A Social Approach*, M. Heller (ed.), 207–230, London: Palgrave Macmillan.
- Androutsopoulos, Jannis. 2013. Code-Switching in Computer Mediated Communication. In *Pragmatics of CMC*, S.C. Herring, D. Stein, T. Virtanen (eds.), 667–694, Berlin/New York: Mouton de Gruyter.
- Anis, Jacques; Christian Duplan. 2001. *Parlez-vous texto? Guide des nouveaux langages du réseau*. Paris: Cherche Midi.
- Anis, Jacques. 2007. Neography — Unconventional spelling in French SMS Text Messages. In *The Multilingual Internet. Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 87–115, Oxford: Oxford University Press.
- Auer, Peter. 1984. *Bilingual Conversation*. Amsterdam: John Benjamins.
- Auer, Peter. 1996. Bilingual Conversation. Dix ans après. *Acquisition et Interaction en Langue Etrangère (AILE)* 7: 9–34.
- Auer, Peter. 2007. Bilinguales Sprechen: (immer noch) eine Herausforderung für die Linguistik. *Sociolinguistica* 20: 1–21.
- Backus, Ad. 2003. Units in codeswitching: Evidence for multimorphemic elements in the lexicon. *Linguistics* 41(1): 83–132.
- Bautista, Maria Lourdes S. 2004. Tagalog-English code switching as a mode of discourse. *Asia Pacific Education Review* 5(2): 226–233.
- Béguelin, Marie-José. 2012. La variation graphique dans le corpus suisse de SMS en français.» In *Penser les langues avec Claire Blanche-Benveniste*, S. Caddéo, M.-N. Roubaud, M. Rouquier, F. Sabio (eds.), 47–63, Aix-en-Provence: Presse de l'Université de Provence.
- Berthele, Raphael. 2004. Vor lauter Linguisten die Sprache nicht mehr sehen. Diglossie und Ideologie in der deutschsprachigen Schweiz. In *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum*, H. Christen (ed.), 111–136, Wien: Edition Praesens.
- Berthele, Raphael. 2010. Investigations into the Folk's Mental Models of Linguistic Varieties. In *Advances in Cognitive Sociolinguistics*, D. Geeraerts, G. Kristiansen, Y. Peirsman (eds.), 265–290, Berlin, New York: de Gruyter.
- Bullock, Barbara E.; Almeida Jacqueline Toribio (eds.). 2009. *The Cambridge Handbook of Linguistic Code-Switching*. Cambridge: Cambridge University Press.
- Carrier, Mark L.; Sandra Y. Benitez. 2010. The effect of bilingualism on communication efficiency in text messages (SMS). *Multilingua* 29(2): 167–183.
- Christen, Helen; Manuela Guntern; Ingrid Hove; Marina Petkova. 2010. *Hochdeutsch in aller Munde. Eine empirische Untersuchung zur gesprochenen Standardsprache in der Deutschschweiz*. Stuttgart: Franz Steiner.

- Crystal, David. 2008. *Txtng. The Gr8 Db8*. Oxford/New York: Oxford University Press.
- Clyne, Michael. 1967. *Transference and Triggering. Observations on the Language Assimilation of Postwar German-Speaking Migrants in Australia*. The Hague: Nijhoff.
- Cougnon, Louise-Amélie. 2011. 'Tu te prends Pour the King of the World?' Language Contact in Text Messaging Context. In *Language Contact in Times of Globalization*, C. Hasselblatt, H.P. Houtzagers, R. van Paren (eds.), 45–59, Amsterdam/New York: Rodopi.
- Chiluwa, Innocent. 2008. Assessing the Nigerianness of SMS text-messages in English. *English Today* 24(01): 51–56.
- Deumert, Ana; Oscar S. Masinyana. 2008. Mobile language choices. The use of English and isiXhosa in .text messages (SMS). *English World-Wide* 29(2): 117–14.
- Dürscheid, Christa; Elisabeth Stark. 2011. SMS4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In *Digital Discourse. Language in the New Media*, C. Thurlow, K. Mroczek (eds.), 299–320, Oxford: Oxford University Press.
- Ellis, Nick C. 2012. Formulaic language and second language acquisition. Zipf and the phrasal teddy bear. In *Annual Review of Applied Linguistics* 32: 17–44.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier. 2006. *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête "Faites don de vos SMS à la science"*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Fillmore, Charles J.; Paul Kay; Mary Katherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64(3): 501–538.
- Frehner, Carmen. 2008. *Email, SMS, MMS. The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern/New York: Peter Lang.
- Gumperz, John J. 1982. *Discourse Strategies*. Cambridge: Cambridge University Press.
- Haggan, Madeline. 2007. Text messaging in Kuwait. Is the medium the message? *Multilingua* 26(4): 427–449.
- Heller, Monica (ed.). 1988. *Codeswitching: Anthropological and Sociolinguistic Perspectives*. Berlin: Mouton de Gruyter.
- Hinnenkamp, Volker. 2008. Deutsch, Doyc or Doitsch? Chatters as language users — The case of a German-Turkish chat room. *International Journal of Multilingualism* 5(3): 253–275.
- Hove, Ingrid. 2008. Zur Unterscheidung des Schweizerdeutschen und der (schweizerischen) Standardsprache. In *Sprechen, Schreiben, Hören. Zur Produktion und Perzeption von Dialekt und Standardsprache zu Beginn des 21. Jahrhunderts*, H. Christen, E. Ziegler (eds.), 63–82, Wien: Edition Praesens.
- Jaffe, Alexandra; Jannis Androutopoulos; Mark Sebba; Sally Johnson (eds.). 2012. *Orthography as Social Action: Scripts, Spelling, Identity and Power*. Berlin: Mouton de Gruyter.
- Jørgensen, Normann J. 2008. Polylingual languaging around and among children and adolescents. *International Journal of Multilingualism* 5(3): 161–176.
- Langacker, Ronald W. 1991. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin/New York: Mouton de Gruyter.
- Li Wei. 2002. 'What do you want me to say?' On the Conversation Analysis approach to bilingual interaction. *Language in Society* 31(2): 159–180.
- Lüdi, Georges; Bernard Py. 2003. *Etre bilingue*. Bern: Peter Lang.
- Lüdi, Georges; Iwar Werlen. 2005. *Sprachenlandschaft in der Schweiz*. Neuchâtel: Bundesamt für Statistik.

- Meeuwis, Michael; Jan Blommaert. 1998. A Monolectal View of Code-Switching: Layered Code-Switching among Zairians in Belgium. In *Code-Switching in Conversation: Language, Interaction and Identity*, P. Auer (ed.), 76–100, London/New York: Routledge.
- Muysken, Pieter. 2000. *Bilingual Speech. A Typology of Code-Mixing*. Cambridge: Cambridge University Press.
- Myers-Scotton, Carol. 1992. Comparing Code Switching and Borrowing. In *Codeswitching*, C.M. Estmann (ed.), 19–39, Exeter: Multilingual Matters.
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, Teddy Arnavielle (ed.), 32–52, Université Paul-Valéry Montpellier 3.
- Pekarek Doehler, Simona. 2011. ‘Hallo! Voulez-vous luncher avec moi hüt?’ — Le code-switching dans la communication par SMS. *Linguistik Online* 48(4).
- Poplack, Shana. 1980. Sometimes I’ll start a sentence in Spanish y termino en español: Towards a typology of code-switching. *Linguistics* 18: 581–618.
- Rash, Felicity. 1998. *The German Language in Switzerland. Multilingualism, Diglossia, and Variation*. Bern: Peter Lang.
- Schwartz, Adam. 2011. Mockery and Appropriation of Spanish in White Spaces: Perceptions of Latinos in the United States. In *The Handbook of Hispanic Sociolinguistics*, M. Díaz-Campos (ed.), 646–663, Oxford: Wiley-Blackwell.
- Sebba, Mark; Anthony Wootton. 1998. We, They and Identity. Sequential vs. Identity-Related Explanation in Codeswitching. In *Code-Switching in Conversation. Language, Interaction and Identity*, P. Auer (ed.), 262–286, London/New York: Routledge.
- Sebba, Mark. 2012. Researching and Theorising Multilingual Texts. In *Language Mixing and Code-Switching in Writing. Approaches to Mixed-Language Written Discourse*, M. Sebba, S. Mahootian, C. Jonsson (eds.), 1–26, New York: Routledge.
- Siebenhaar, Beat. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics* 10(4): 481–506.
- Spilioti, Tereza. 2009. Graphemic representation of text-messaging: Alphabet-choice and code-switches in Greek SMS. *Pragmatics* 19(3): 393–412.
- Spycher, Samuel. 2004. “I schribdr de no...”. Schweizerdeutsche Umgangsformen in der SMS-Kommunikation. *Networx* 36.
- Tagg, Caroline. 2009. *A Corpus Linguistics Study of SMS Text Messaging*. Unpublished Ph.D. dissertation French (Switzerland) at the Department of English, University of Birmingham.
- Thibault, André; Pierre Knecht. 2004. *Dictionnaire suisse romand: particularités lexicales du français contemporain*. Carouge-Genève: Editions Zoé.
- Thurlow, Crispin; Michele Poff. 2013. Text Messaging. In *Pragmatics of CMC*, S.C. Herring, D. Stein, T. Virtanen (eds.), 163–190, Berlin/New York: Mouton de Gruyter.
- Winter-Froemel, Esme. 2011. *Entlehnung in der Kommunikation und im Sprachwandel. Theorie und Analysen zum Französischen*. Berlin: Mouton de Gruyter.
- Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Online References

<http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/c#mobilemessaging> (retrieved on 21 August 2012)

<http://www.sms4science.org> (retrieved on 14 Februari 2013)
<http://www.sms4science.ch> (retrieved on 14 Februari 2013)
<http://www.tagesanzeiger.ch/wirtschaft/unternehmen-und-konjunktur/GratisSMSProgramme-setzen-den-Schweizer-Telecomfirmen-zu/story/31419158> (retrieved on 14 Februari 2013)
http://wortschatz.uni-leipzig.de/cgi-portal/de/wort_www?site=208&Wort_id=17206610 (retrieved on 14 Februari 2013)
<http://www.youtube.com/watch?v=HdQOaHr9Ltg> (retrieved on 14 Februari 2013)

Summary

The use of more than one language in SMS communication is widespread, yet has remained relatively underexplored in the existing research. In this paper we ask: What methodological and conceptual tools are needed for empirically investigating code-switching in large databases of SMS communication? We show that the investigation of SMS communication calls for an adaptation of the conceptual and the methodological apparatus of classical code-switching studies, which have been typically concerned with the analysis of spoken, mostly interactional, data. We argue for a broad understanding of code-switching that comprises switching between natural languages and language varieties along with style shifts as well as switching between language and other semiotic systems (ideographic switching). We also document, as a key feature of SMS communication, hybrid forms of language use that blur the boundaries between what we commonly call languages (e.g. homographs, mixed spellings or allogenisms), and we suggest that these possibly indicate that SMS communication has become one site where the tension between localized and globalized social practices is played out. The study presented here is part of an inter-university research project, entitled “SMS communication in Switzerland: Facets of linguistic variation in a multilingual country”, based on a corpus of 26,000 authentic messages collected between 2009 and 2011.

French text messages

From SMS data collection to preliminary analysis*

Rachel Panckhurst* and Claudine Moïse**

*Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3 /

**Lidilem, Université Stendhal Grenoble 3

Introduction

Over a three-month period (spanning 15 September to 15 December 2011), over 90,000 authentic text messages in French were collected by a group of academics in the Languedoc-Roussillon region of France, in the context of the following programme: *sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée*.¹ Participants from the general public (both children/teenagers and adults) were invited to enrol online and give their consent

* The authors would like to warmly thank their colleagues and doctoral student for the fruitful collaboration during the *sud4science LR* project: Catherine Détrie, Cédric Lopez, Mathieu Roche, Bertrand Verine. We would also like to thank three M1 and two M2 students who worked on the project via internships (Antony Stifani, Namrata Patel, Pierre Accorsi, Michel Otell, Camille Lagarde-Belleville). Several of our *sud4science LR* colleagues and students, Dr Fay Panckhurst, and two anonymous reviewers made constructive remarks on a previous version of this paper and we thank them for these. The CENTAL team at UCL provided significant help for the data collection (see note 13) for which we are very grateful.

1. This programme took place from January 2011 to December 2012 within the *Maison des Sciences de l'Homme de Montpellier* (Montpellier Human Sciences House — MSH-M). Data collection (<http://www.sud4science.org>), cleanup and de-identification/anonymisation of the SMS corpus were conducted (<http://www.msh-m.fr/programmes-2012/sud4science-lr/>). Video presentations of the scientific seminars which were also organised during the two-year period are available at: <http://www.msh-m.tv/spip.php?rubrique138>. The *sud4science LR* programme is part of a vast international SMS data collection project, entitled *sms4science* (<http://www.sms4science.org>, Fairon, Klein, Paumier, 2006), and started at the CENTAL (*Centre for Natural Language Processing*, Université Catholique de Louvain, Belgium) in 2004. A second research project began while the authors were finalising this article: *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS*, (Délégation générale à la langue française et aux langues de France, DGLFLF) and it runs from 1/7/12 to 30/6/13. It will enable the researchers to move from the *anonymisation* stage to the *transcoding*

(<http://www.sud4science.org>), before donating their own authentic text messages. Two possibilities for donating the messages were used: either the participants forwarded messages to the research team which they had already sent to others and which had been previously saved in the memory of their mobile phone; or, when the participants were on the verge of sending a text message to others, they could simultaneously send a copy of it to the researchers. Participants were also invited to fill out a sociolinguistic questionnaire.

For this contribution, the authors decided to outline research processes rather than provide in-depth analytical results. It provides insight into behind-the-scenes work and trial-and-error aspects of the project (B. Matalon 1996), which are often significantly time-consuming and full of questions, but nevertheless essential parts of research. Heuristics are important, and this approach was adopted in order to avoid taking the tools and the collected data for granted.

This chapter retraces the organisation of the data collection (§ 1), the elaboration of the sociolinguistic questionnaire that donors were invited to fill out (§ 2), and text message data processing procedures and preliminary results (§ 3). A shift from individual “isolated” text messages to “conversational” SMS exchanges is then studied, in preparation for a new SMS conversational data collection which is due to take place in the near future (§ 4). This whole process is important for understanding in-depth interactional practices within contemporary digital textuality and should provide insight for pluridisciplinary approaches.

1. Data collection: 93,085² text messages

508³ people from the general public enrolled to participate on the project website (<http://www.sud4science.org>). An optional sociolinguistic questionnaire was also proposed, and 95% chose to complete this. Of the 483 people who filled out both the consent form (compulsory in order to subsequently donate text messages) and the questionnaire (compulsory only for those who chose to take part in a weekly

stage of the *sud4science* LR corpus before conducting analysis of the text message data and the related sociolinguistic questionnaire.

2. Depending on the software used, the total figure varies from 93,114 (Phoneview) to 93,085 (Microsoft Excel, Open Office). This may also be due to character recognition (no of characters for accents, UTF-8, etc.) The total number of characters/words is as follows: total number of words: 1,252,057; total number of characters with spaces: 6,177,191; total number of characters without spaces: 5,022,787. See note 7 for statistics related to “cleansed” corpus.

3. 509 people enrolled on the website, but one of these was a hoax (age 99, etc.)

lottery⁴), 435 actually became text message donors.⁵ Given that 93,085 messages were collected during the 13-week period, each donor sent 214 messages on average. Since some people sent as few as 5 messages (the minimum required in order to participate in the weekly lottery), this average is very high.⁶ Compared to the first SMS data collection, conducted in Belgium in 2004 by Cédric Fairon and his team from the CENTAL, Université Catholique de Louvain (C. Fairon *et al* 2006), these statistics are quite different (*cf.* Table 1 below).

One of the obvious reasons which explains this significant difference is that a very high percentage of mobile phone monthly plans in 2004 would not have included unlimited text messages. In recent years, this has radically changed and 88% of the 2011 *sud4science LR* donors have this feature included in their monthly plans. Interestingly, probably owing to this fact, some teenagers (have indicated to the authors in private conversations that they) now use the unlimited text message feature as an alternative to online written chats, given they always have their phones with them. Another reason for the high average is possibly that a prize was

4. A weekly lottery was organised in order to thank participants for having donated their text messages. Every Thursday, from 22 September to 15 December 2011, 4 or 5 donors were drawn and prizes sent to them (after enrolment & consent verification). Among these were: an iPod shuffle, books & vouchers, a haircut, chocolate CDs, a beauty treatment, cooking classes, film tickets. At the end of the three-month period, three prizes were issued to the donors who sent in the most text messages. Prize winner number 1, a female fourth year University student aged 21 sent in 8,045 text messages in 10 weeks — she won an iPad 2 16GB, donated by our commercial computer partner (*iTribu* <http://www.itribustore.fr/>); prize winner number two, a female second year student aged 22 donated 6,540 text messages — she won an oil change (*Carrosserie Guignard*); prize winner number 3, also a female third year student aged 20 donated 3,900 text messages — she won a meal for two (*Imoë* <http://imoerestaurant.canalblog.com/>). A fourth year Montpellier 3 University student, Anthony Stifani, was employed for a three-month internship, in order to read the incoming messages and make sure they respected our rules and regulations. A doctoral student working on our project, Cédric Lopez, designed and implemented the computer program which allowed us to select the weekly winners using computer random methods.

5. Of the 435 donors, 11 participants did not fill out the consent form, therefore their text messages were eliminated from the final corpus. The final figure is therefore 424 donors. This means that 59 people either chose not to send any text messages after having enrolled and filled out the questionnaire, or that there was a technical problem which meant that their text messages were not received by the research team. If one includes those participants who solely filled out the consent form (and not the questionnaire) and never sent text messages during the three-month period, that figure increases to 84. The term ‘donor’ is retained here, instead of ‘contributor’ in reference to the original medical ‘pun’ used in the Université Catholique de Louvain’s 2004 project: “Faites don de vos SMS à la science”.

6. The top ten participants sent a total of 37,133 SMS, *i.e.* 40% of the total corpus. The average for these donors is therefore 3,713.3 messages per person, which is exceedingly high and of course contributes in increasing the general average.

Table 1. Donor/text message ratio

	No of donors	Average No of text messages per donor	Total No of text messages ⁷
<i>smspurlascience</i> (Louvain, 2004)	3,600	21	75,000
<i>sud4science LR</i> (Montpellier, 2011)	435	214	93,085
<i>sud4science LR</i> ("cleansed" corpus)	424	209	88,683

issued to each of the three donors who had sent in the most messages, including an iPad, for winner No 1, which may have been an important incentive. The iPad winner sent in 8,045 SMSes⁸ over a 10-week period, and the top ten contributors account for 40% of the overall messages received.

Several SMS data collections, within the overall framework of the *sms4science* project, have taken place since the initial Belgian one in 2004 (Reunion Island (L.-A. Cougnon & G. Ledegen 2010), Switzerland (C. Dürscheid & E. Stark 2011), Quebec (Langlais *et al.* 2012), French Rhône-Alps (Antoniadis *et al.* 2011)).⁹ In

7. The Université Catholique de Louvain corpus was then “cleansed” and a total of 30,000 messages were used in the final corpus. The *sud4science LR* number of text messages decreased from 93,085 to 88,683 after elimination of spurious messages. Cf. § 3.2 for explanations (4,402 messages were suppressed: 4,123 SMS were received twice or more from the same participants; 113 SMSes were sent by Rachel Panckhurst to participants; 54 SMSes were received from telephone companies and another 54 SMSes from abroad; 40 messages were received from those who had not signed the consent form; 18 advertisement SMSes were received). The statistics for characters/words are as follows: total number of words: 1,218,749 (Microsoft Word), 1,219,771 (OpenOffice); total number of characters with spaces: 5,984,306 (Word & OpenOffice); total number of characters without spaces: 4,858,902 (Word); number of characters with spaces on average per SMS: 67.47 (Word, OpenOffice); number of characters without spaces on average per SMS: 54.78 (Word, OpenOffice); number of words on average per SMS: 13.74 (Word), 13.75 (OpenOffice). See note iii for statistics related to original corpus. The statistics for the other *sms4science* data collections are as follows: 2,514,399 characters, 688,550 tokens (Belgium); 748,484 characters, 233,285 tokens (Reunion Island); 225,256 characters, 60,809 tokens (Quebec); 374,644 characters, 94,398 tokens (Switzerland); 1,292,544 characters, 239,290 tokens (Rhône-Alps, France). See the introduction to this special issue for detailed statistics.

8. In a conversation with the iPad winner in June 2012, we questioned the average amount of messages sent during the data collection and the person confirmed that this was a higher average than her usual amount of text message communication, owing to having also donated SMS that she had already sent and which were present in the phone memory.

9. In chronological order, after the initial Belgian collection, the following regions & countries collected authentic text messages, within the framework of the international *sms4science* project:

terms of data-collection methods, *sud4science LR* used a novel approach. Instead of having donors forward/copy their text messages to a short four-digit telephone number (used frequently in France for donations from the general public and organised by one of the telephone companies who then forwards the messages to the research team), the Montpellier researchers decided to use an iPhone for the data collection.¹⁰ Once the donors had completed the online consent and enrolment, they were then given a normal ten-digit mobile phone number and invited to send as many text messages as they liked for the 13-week period. The fact that the process was made easy for donors¹¹ most likely contributed to the success rate in terms of the high amount of messages received. The team was hoping to receive 50,000 SMSes and they received an overwhelming 93,085 in just 13 weeks.

Once the messages were received on the iPhone, each week they were transferred to an offline computer via a piece of software (*Phoneview*, <http://www.ecamm.com/mac/phoneview/>). This tool meant that the data were easily extractable and copied into various formats (text, portable document format (PDF), comma separated values (csv) and database), ready for subsequent data processing (thus also avoiding synchronising on the computer). An internship was allocated to a graduate student during the data collection period in order to read/skim through the messages and make sure they corresponded to French privacy rules and regulations (see § 3.2. and note 17).

During this period, a doctoral student specialising in computer science (*cf.* note 4) designed a computer program in order to i) select random lottery winners, (ii) check that donors had in fact enrolled and filled out the online consent, (iii) stipulate which donors were among the 'top ten' contributors (in terms of numbers of text messages received) of the week.

Reunion Island: <http://www.lareunion4science.org/> (20,000 SMSes, 2008); Switzerland: <http://www.sms4science.ch/> (24,000 SMSes, 2009–2010); Québec: <http://www.texto4science.ca/> (5,000 SMSes, 2010); Grenoble: <http://www.alpes4science.org/> (22,000 SMSes, 2010).

10. We contacted a commercial computer partner in Montpellier, *iTribu* (<http://www.itribustore.fr/>), and we received a warm welcome; the company's directors, Patrice & Christell Belot, accepted to be associated with our programme and lent us an iPhone with a monthly plan including unlimited text messages for the duration of our collection.

11. One of the drawbacks in some of the previous data collections was that donors had to indicate a code at the beginning of their donated text message, because the short four-digit numbers were shared between different projects. If a donor forgot to do so, or if the text message was longer than 160 characters (and therefore divided into two or more messages), then the message was lost or truncated because there was no code at the beginning of the second (and subsequent) part(s) of the SMS.

Concerning the number of characters, the messages are on average 55 characters excluding spaces, and 67 characters including spaces. Each message therefore contains, on average, 13.75 words. After the transcoding stage of the data processing, during which text messages are transcribed into ‘standardised’ French (see § 3.2.2), it will be interesting to check the abbreviation rate between “raw” unprocessed text messages and transcribed ones.¹²

2. Linking data and usage: Sociolinguistic questionnaire

In order to improve our understanding of SMS usage, researching authentic data is essential. For this reason, in 2004, the Belgian CENTAL research team decided to gather a global database in order to study mediated discourse in text messages. Since SMSes are often linked with neological/neographical “creativity” (among others, L.-A. Cougnon, 2012, C. Fairon *et al.* 2006, R. Panckhurst 2009), it is important to conduct research related to usage, in order to understand how “creativity” is perceived and how it may evolve. For instance, with smartphone usage of predictive text input, are text messages now longer (compared to 2004)? Have mutations taken place with the shift (in France at least) from limited to unlimited text message monthly plans? Collecting authentic text messages (and not artificially created ones) from the general public is therefore essential.

Comparing text message usage with other types of electronic discourse (e-mail, forums, chats), normative writing on paper and oral exchanges is also important. This allows us to observe different types of writing, choice of vocabulary, sentence construction, etc. These factors vary depending on people, age, gender and situations. For instance, are openings used (“cc”, “bjr”, “slt” — “hi”, “hello”, “giddy”, etc.) or does the texter launch straight in to the text message? How does the texter address the recipient: by using a first name, a nickname, a tender word, or other forms? Have these evolved over the past ten years?

12. Text messages are not always abbreviated, because they often contain extensions, including character repetition, etc. The abbreviation rates vary from project to project: L.-A. Cougnon & T. François (2011) indicate that the Belgian 2004 study includes a 9.5% abbreviation rate, compared to 17.5% (Reunion Island) and 14.6% (Québec). The Belgian study also indicates a 17.3% usage of very short text messages (≤ 15 characters). In the 2011 *sud4science LR* corpus, the number of very short messages (≤ 15 characters) corresponds to 14,677 SMSes, i.e. 16.55% of the overall “cleansed” corpus. It is interesting to compare this with the 2004 Belgian study, in which the percentage is slightly higher; on the contrary, one might have suspected the percentage of very short messages to be superior in the 2011 corpus, since monthly plans often include unlimited text message features and texters no longer hesitate to send very short messages such as “OK” in response to text messages they receive.

Authentic text messages are ideally studied along with information about telephone and text message usage. A sociolinguistic questionnaire (*cf.* Appendix), based on previous SMS data collections, was compiled in order to gather information of two kinds: (1) *practices* and *usages*; (2) *representations* related to these usages. Given the age range of participants, clarity and simplicity of the questions was essential. Both open and closed questions were used. Ethical preoccupations, related to private data collection were also considered, and important work was conducted with the legal services of the University in order to provide a coherent enrolment and legal consent form as well as clear and structured information about the aims of the project to participants. Technical aspects were also highly important.

The questionnaire process involves the choice of questions, then having the questionnaire filled out by participants, before coding and examining data. Epistemological issues are therefore crucial, because the way in which questions are formulated can cause misinterpretation and thus shape a certain perception of reality.

Mistakes can be due to lexical ambivalency, for instance: “quelle(s) langue(s) parlez-vous couramment” (*what language(s) do you speak fluently?*). In French, “couramment” could mean either “often” or “fluently” and so the authors changed their initial wording of the question to “Dans quelle(s) langue(s) vous sentez-vous le plus à l’aise ?” (*In which language(s) do you feel the most at ease?*) as it seemed more explicit. A question which is too long is sometimes more difficult to understand than a concise one: “Pensez-vous qu’à l’avenir, notamment avec l’utilisation des téléphones intelligents, ou avec les appels illimités vers les mobiles depuis un téléphone fixe, ou encore avec des outils permettant de téléphoner gratuitement en zone wifi (par exemple Skype, Viber, Facetime...) que les SMS seront encore utilisés ? Merci d’expliquer votre réponse, en nous donnant, si possible, des exemples.” (*Do you think that in the future, for instance with smartphone usage, or unlimited landline to mobile calls, or tools which allow free calls in wifi zones (Skype, Viber, Facetime...) that text messages will still be used? Please highlight your answer, if possible, with examples.*) The authors changed this to: “Pensez-vous qu’à l’avenir les SMS seront encore utilisés ? Merci d’expliquer votre réponse, en nous donnant, si possible, des exemples.” (*Do you think that text messages will be used in the future? Please highlight your answer, if possible, with examples.*)

Questions can also provoke subjective interpretation. Following a period of discussion and thinking through issues between researchers, the authors decided to change another question which was seemingly too positive (and therefore subjective) “D’après-vous, qu’est-ce qu’apporte l’écriture SMS ?” (*According to you, what does SMS writing provide?*) to another more neutral one “D’après-vous, quels sont les avantages/inconvénients de l’écriture SMS ?” (*According to you, what are the advantages/disadvantages of SMS writing?*)

2.1 Preliminary work and pluri-disciplinary skills

Several researchers of the *sud4science LR* group contributed to developing the questionnaire. The Belgian CENTAL team also contributed significantly during preliminary phases, with matters related to the questionnaire, data collection phases, etc.¹³ In order to improve the technical online functioning of the questionnaire, the *sud4science LR* project leader and one of the CENTAL computer scientists communicated daily with sustained exchanges, concerning, among other issues:

- *how to respond to a user having already filled out a questionnaire, if it is done a second time;*
- *making sure a user types his/her telephone number twice, in order to reduce the possibility of typing mistakes;*
- *practical questions concerning a user-friendly interface (back buttons within the template pages of the questionnaire, as traditional browser buttons were deactivated whilst filling out the questionnaire; a progress toolbar at the top of the page so the user could check on the approximate amount of time needed to complete the questionnaire, etc.)*

Owing to effective and extensive collaboration between all research partners, and given that the Montpellier data collection took place after many others (see introduction to this special issue), the researchers also benefited from previous ‘trial and error’ situations. The final result was therefore relatively easy to manipulate for participants of all ages.

2.2 Sociological aspects

The way in which the questionnaire is drafted sheds light on sociological aspects, including gender categories. At a time when gender studies strive to discard binary representations between women and men, the authors chose to enable a checkbox click on both “male” and “female” genders, if a user so wished, in order to avoid gender distinction. In actual fact, besides one hoax answer, all respondents chose one or the other gender. Similarly, in-depth discussions between researchers were conducted related to professional status, since the way in which social categories are mentioned can provoke social distinction. The researchers therefore decided to indicate general professional fields and remain as open as possible. The users

13. Louise-Amélie Coughon spent 70 hours, exchanged 300 emails, and had two telephone conversations and two online video chats helping the Montpellier researchers with logistics, media aspects and legal matters. Hubert Naets spent 250 hours dealing with technical issues related to the *sud4science.org* website and the questionnaire.

could tick several boxes, thus combining several categories belonging to different fields, such as “health profession” and “senior position”. The “other” box offered a wide range of answers: 39 indications described more specific jobs like “qualified childminder” or “web developer”, etc. (see Figure 3 for overall professional fields).

2.3 Sociolinguistic aspects

From a sociolinguistic point of view, the questionnaire focuses on two points, languages used and representations related to SMS writing.

2.3.1 *Language choice*

The name of a language is very important for defining what it represents for its speakers. The researchers therefore spent a long time discussing language choice. What kind of languages were to be promoted? According to which linguistic ideology? Should Occitan and Provençal be distinguished? Should languages appearing in the European Charter for Regional or Minority languages (such as Yiddish and Wolof) be given priority? Or those from the South of France? Several interesting exchanges were made between researchers concerning language options: “We could indicate the following languages: 1. French, 2. Arabic (different countries), 3. Occitan, 4. French Sign Language, 5. Other(s): indicate which language(s); or, if we want to include social categories from the south, we could mention: 1. French, 2. Arabic, 3. Spanish, 4. Berber, 5. Occitan, 6. Catalan, 7. French Sign Language, 8. Other(s)”. After much brainstorming, the ISO-639-2 code, (a set of international standards listing short codes for language names) including 463 languages was chosen so that users would not be influenced by a particular choice the researchers might have made, had they decided to list a limited number of languages. However, the computer scientist then indicated the technical difficulties linked to this choice, i.e., that the list would need to be loaded 12 times on the “general” page, and it might therefore be a better idea to suggest the most “likely” languages at the top of the list. The research group responded negatively to this request, so as to not ‘influence’ the user’s choice, from a sociolinguistic point of view. However, they specified that certain languages should not be distinguished because of important linguistic debates (e.g. Occitan and Provençal) but that others could be sub-divided: “Local Arab Dialect (Moroccan, Algerian, Tunisian...)”.

2.3.2 *Languages used*

Contrary to common practice, the researchers decided not to indicate the question related to native languages first, but rather the one concerning “language(s) one feels at ease with”. Otherwise, they felt that answers could have been influenced. The following three questions were presented, respectively: 1. “Dans quelle(s)

langue(s) vous sentez-vous le plus à l'aise ?"», 2. "Quelle(s) langue(s) employez-vous le plus souvent ?", 3. "Quelle(s) est/sont votre/vos langue(s) maternelle(s) ?" (1. *In which language(s) do you feel the most at ease?*, 2. *Which language(s) do you use the most often?*, 3. *What is/are your native tongue(s)?*).

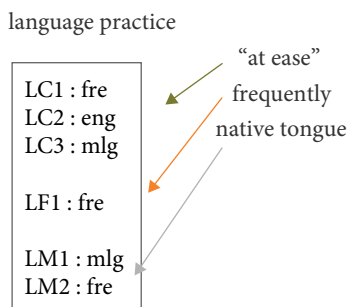


Figure 1. An example of a participant's language use
(LC=*langue courante*, LF=*langue fréquente*, LM=*langue maternelle*)

To highlight this, in Figure 1, the participant's native languages are Malagasy and French, respectively, but (s)he feels most at ease using French, English and Malagasy in that order, and uses French most often on a day-to-day basis (see Appendix for full questions as they appear to users and § 3.1.).

2.3.3 Representations of SMS writing

Unlike other *sms4science* studies the researchers wanted to focus on how people perceive *SMS writing* (see R. Panckhurst 2009; L.-A. Cougnon & T. François 2011; L.-A. Cougnon 2012 for definitions of “écriture/écrit SMS” “*SMS writing*”, “*SMS script*”, terms which the authors prefer compared to “textisms”, “textese”, “text-speak”, etc.) but without overloading the questionnaire and thus discouraging participants. This section was as open as possible, yet provided avenues for possible answers. Two questions were provided, as follows: 1. “Pourquoi écrivez-vous des SMS (au lieu de téléphoner, par exemple) ?” (1. *Why do you use text messaging (instead of telephoning, for instance)?*) with possible checkbox answers: “parce que c’est moins cher ou inclus dans le forfait” (*because it’s cheaper or included in the monthly plan*), “pour ne pas déranger” (*so as to not disturb people*), “pour aller plus vite” (*to be quicker*), “parce que je n’aime pas téléphoner” (*because I don’t like to phone*), and an “Other” open final answer; 2. An optional question: “Si vous rédigez en écriture SMS, pourquoi le faites-vous ?” (2. *If you use SMS writing, why do you do so?*) The following answers were associated with this question: “parce que c’est plus rapide” (*because it’s quicker*), “parce que j’aime jouer avec la langue” (*because I like playing with language*) “parce que ça crée une complicité avec les amis” (*because it creates a close bond with friends*) and “Other”.

This was the only category where the researchers provided possible open answers,¹⁴ so that participants could indicate their own propositions and interpretations, which they in fact did do. The next stage will be the in-depth analysis of all of the answers provided within the sociolinguistic questionnaire.

3. Preliminary results and data processing procedures

3.1 Participants, professions, languages

A majority of the participants are young: 80% are under 30. More women participated than men (62% vs. 38%), and the child/teenager vs. adult ratio is as follows: 29% are under 18 vs. 71% who are 18 and over. The age range spans from 11 (first year of secondary school in France) to 66 and the average age is 24. Sixty-five percent of the participants have completed or are enrolled at a tertiary education institution and, among these, 63% are currently undergraduate students. A high ratio (80%) have been using text messages for over five years (75% of these are under 30).

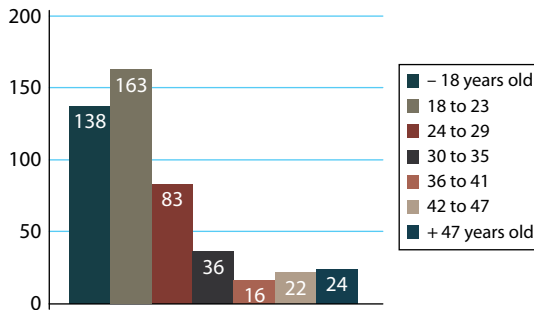


Figure 2. No of participants per age span

An interesting fact is that 74% of the participants have a smartphone (*iPhone*: 53%; *Blackberry*: 21%). Given that the study started out in a Human Sciences University with a large proportion of government bursars whose parents have low incomes, it is clear that a modern mobile telephone is a top priority for the young, or even a necessity.

14. The authors suspect there may have been a misinterpretation of the question concerning SMS writing, since many people in the general public refer to *langage SMS* (SMS language) rather than *écriture SMS* (SMS writing). The second question actually refers to whether participants like to abbreviate words, extend characters, and generally 'play' with spelling in texts, etc. However, given some of the answers in the questionnaire, this may have been understood as using texting in general. A full analysis of the questionnaire in due course will enable the authors to clarify this.

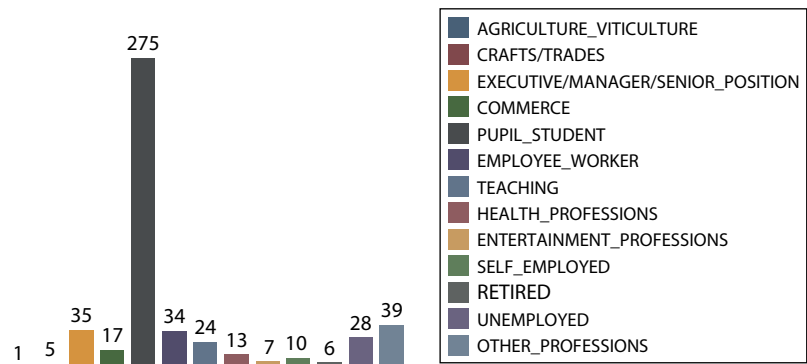


Figure 3. Professional fields of questionnaire respondents in numbers

Thirty-two languages were quoted by users.¹⁵ Of these, Table 2 gives indications on the four most frequent first position languages: “at ease” (LC), most frequently used (LF) and native tongues (LM).

Table 2. Languages appearing in first position by decreasing frequency

LC1	LF1	LM1
fre	fre	fre
eng	eng	eng
spa	spa	wol
ita	ara_d, bam, ger, ita, sgn, srp, vie, wol	ara_d, ger, mlg, spa

In the left-hand column, first-position “at ease” (LC) languages indicated are French, English, Spanish and Italian, respectively. These do not necessarily coincide with frequently-used languages and native tongues; for instance, Wolof is indicated as a more frequent (than Spanish) first position native tongue for some users. User languages will be compared with SMS data in order to study ‘code-switching’ and other language-related issues in future research.

15. The 32 languages, in alphabetical order, are: ara_c (classical Arabic), ara_d (dialectal Arabic), bam(Bambara), ber(Berber languages), cat(Catalan), chi(Chinese), cpf (Creoles and pidgins, French-based), cpp (Creoles and pidgins, Portuguese-based), eng(English), fre(French), ger(German), gre(Greek), heb(Hebrew), ita(Italian) jpn (Japanese), kab(Kabyle), kaz(Kazakh), lat(Latin), mao(Maori), mlg (Malagasy), oci (Occitan), pol(Polish), por(Portuguese), rum (Romanian; Moldavian), rus(Russian), sgn (French sign language), spa(Spanish; Castilian), srp (Serbian), swe(Swedish), tur(Turkish), vie(Vietnamese), wol(Wolof).

3.2 Data processing procedures

During a two-month period (June–July 2012) two computer science university students¹⁶ completed an internship in order to anonymise the *sud4science LR* corpus. The computer program they devised for SMS anonymisation works in three phases: (1) automatic phase: automatic processing of the corpus and labelling of information which needs to be anonymised; (2) semi-automatic phase: processing of remaining items of the corpus which are not possible to treat automatically, via a web interface; (3) validation/approval phase: correction of mistakes. A preliminary phase was also conducted in order to eliminate spurious information (text messages received twice or more from the same person; those received from abroad (since the study was confined to mainland France) or sent by the *sud4science LR* iphone to participants; automatic text messages received from telephone operators, etc.). Anonymisation is a legal requirement given that text messages contain personal, therefore potentially sensitive, information.¹⁷

3.2.1 Anonymisation software

As specified above, the text messages are anonymised via a computer program in a three-phase process (see ‘Seek&Hide’, P. Accorsi *et al.* 2012 in this volume for full details, including a list of all anonymisation tags used): (i) automatic; (ii) semi-automatic; (iii) correction of any mistakes. During phase (i), a dictionary of first names is used in order to recognise, extract and anonymise the first names which appear in the corpus. Several problems need to be addressed: (a) capitalisation (or not) of first names; (b) specific lexical issues, such as character repetition (*Cééééédric*) and diminutive/abbreviated forms (*Nico* for *Nicolas*). Phase ii) requires a computer program/interface to be devised in order to help the encoder visualise candidates which need to be anonymised (first names, places, telephone numbers, various codes, etc.). This identification also includes checking that the word encountered in the first names dictionary and which is to be anonymised does not also appear in an electronic inflected forms dictionary. For instance, a proper name may conflict with an item appearing in the electronic dictionary and provoke ambiguity which needs to be resolved through human processing at the

16. These two Masters’ students Pierre Accorsi and Namrata Patel (LIRMM, Université Montpellier 2), were chosen to do an internship for this purpose.

17. Our University’s *National Commission for Data Protection and Liberties* (CNIL) correspondent, Nicolas Hvoinsky, has made a declaration rendering the anonymisation operation compulsory by 30 September 2013. The CNIL (<http://www.cnil.fr/english/>) indicates the following motto on its homepage: “Information technology must respect the human identity, the human rights, privacy and liberties.”

web interface stage: “[...] ya des gens beaucoup mieu placer que moi pour te comprendre et *ben* je peux essaye [...]”.¹⁸ A number of words may be injected into phase ii) owing to the fact that they do not appear in either the first names dictionary or the electronic inflected forms dictionary. These will not necessarily be subsequently anonymised. A lexicon of frequent SMS abbreviations was therefore also used, but it is of course highly imperfect owing to the extensive spelling creativity appearing in text messages.¹⁹

An example of completed anonymisation for first names appears in Table 3:

Table 3. From ‘raw’ to anonymous SMS

Coco est pas la ! Éva non plus ! Tanpis ! Lol J’irai aux journée du patrimoine ! Éva m’a dit que tu venais cette semaine peut etre ! Bisous ! !
<PRE_4> est pas la ! <PRE_3> non plus ! Tanpis ! Lol J’irai aux journée du patrimoine ! <PRE_3> m’a dit que tu venais cette semaine peut etre ! Bisous ! ! <i>Translation : <First_name_4_characters> isn’t there ! Neither is <First_name_3_characters> ! Too bad! I’m going to the Heritage Day! <First_name_3_characters> said you might be coming this week! Kisses!!</i>

Numbers refer to characters in anonymised first name (PRE=*prénom*, in French)

3.2.2 Transcoding

The transcoding phase allows the text message to be converted into “standardised” French. The term ‘transcoding’ has been chosen for SMS script instead of the more traditional ‘transcription’, which is often used for oral/audio data. However, the original anonymous SMS is of course kept for subsequent linguistic analysis purposes, and no additional interpretations may be included during this phase. For instance, in the following example, only lexical, gramatical or typographical modifications are indicated, but for instance, part of an ellipsis — ‘missing’ negation (“*ne*” in Table 4) is not, as this would indeed modify the corpus by introducing an entirely new element. The idea is to rectify spelling and grammar, yet not ‘inject’ more than is necessary for clear understanding. Both the original text message and the transcoded one will be available after the process has been completed. This

18. “There are some people who are in a better position to understand you and *well*, I can try”. *Ben* is ambiguous in French between the first name and the interjection “*Bien*” abbreviated in this context to “*Ben*”. The authors thank Namrata Patel & Pierre Accorsi for providing this example.

19. Among these, words with spelling mistakes (*surment*, most likely), no accents (*desole*, sorry), accents incorrectly positioned or placed (*dèsolè*, sorry), character repetition (*nicooooolllaasss*, Nicolas), onomatopoeia (*mouhahaha*, hahaha), elision without apostrophes (*jexplique*, I explain), agglutination (*jtaime*, I love you). The authors thank Namrata Patel & Pierre Accorsi for these examples.

phase is useful for the general public, or those working on SMS analysis and who want to be able to read and compare “raw” and/or processed SMS quickly.

Table 4. From anonymous to transcoded SMS (indicated in bold)

<PRE_4> est pas là ! <PRE_3> non plus ! Tanpis ! Lol J’irai aux journées du patrimoine !
<PRE_3> m’a dit que tu venais cette semaine peut etre ! Bisous !!

<PRE_4> est pas là ! <PRE_3> non plus ! Tant pis ! Lol. J’irai aux journées du patrimoine !
<PRE_3> m’a dit que tu venais cette semaine peut-être ! Bisous !!

3.2.3 Annotation

A total of eight tags will be used for this phase: ABSence, BINettes (smileys), DIVers (diverse), GRAMmaire (grammar), LANgage (language), MODification, ORThographe (spelling), TYPographie (typography). Usage of three of these tags is indicated in the example in Table 5 below (note that more than one tag can be used in the case of ambiguity):

Table 5. From transcoded to annotated SMS (tags indicated in bold)

<PRE_4> est pas là ! <PRE_3> non plus ! Tant pis ! Lol. J’irai aux journées du patrimoine !
<PRE_3> m’a dit que tu venais cette semaine peut-être ! Bisous !!

<PRE_4> <ABS_n’> est pas <TYP_’>là ! <PRE_3> non plus ! <MOD_t_TYP_space>Tant pis ! <MOD_laughing out loud>Lol <TYP_>. J’irai aux <GRA_MOD>journées du patrimoine ! <PRE_3> m’a dit que tu venais cette semaine <TYP_-^>peut-être ! Bisous !!

The choice of tags evolved from in-depth discussions with previous organisers of *sms4science* researchers and students conducting annotation on SMS corpora — tags need to provide enough general information, yet not be too ‘noisy’ or ‘silent’ from a documentary point of view, during the annotation phase. Annotation is an optional phase, but if it is conducted, it would allow researchers, teachers, people from varying scientific backgrounds to (semi-) automatically extract specific information in relation to their specialities.

3.2.4 Linguistic analysis

Linguistic data analysis has recently commenced, including lexical, syntactic, semantic and pragmatic phases. C. Détrie & B. Verine (2012) have focused on insults which are toned down or even suppressed by inclusion of tender words and ‘sweet nothings’, which are used when addressing recipients, and are sometimes highly creative. R. Panckhurst and C. Moïse (2011, 2012) have started research on “conversational” text messages (*cf.* § 4).

Concordancing programmes²⁰ allowing for lexical searches indicate that the highest (1-gram) word count in our corpus is allocated to the first person subject pronoun “je” (“I”, 36,153 tokens); the highest 2-gram word count is: “c est” (“It is”, 12,401 tokens, without the usual apostrophe); the most common 3-gram word count is “je t aime”²¹ (“I love you”, 3,110 tokens, also without an apostrophe), followed by “je sais pas” (“I don’t know”, 1,414 tokens, without part of the negative particule “ne”), then “c’est pas” (“it’s not”, 1,244 tokens, also without the complete negative form “ne”, used in standard French). More thorough indications will be given in future publications, once anonymisation has finished.

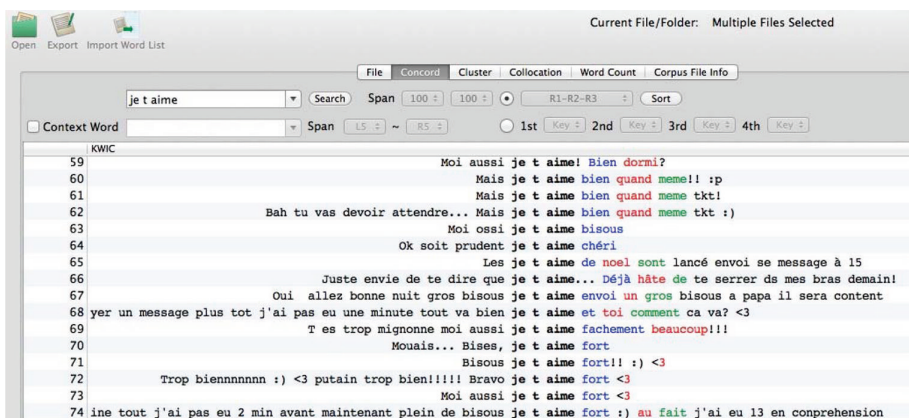


Figure 4. Key words in context (KWIC) “je t aime”, *sud4science LR* corpus, using *Casual Conc*

One might have expected smiley usage to be essentially graphical (😊😊😊😊), rather than with character usage (“:”), “^^”, “:p”, “<3”, etc.), given that certain telephones now permit insertion of a range of graphical smileys. However, this is not the case in the *sud4science LR* corpus. Of the just under 30,000 smiley tokens used, only 1% are graphical.

An isolated lexical search is obviously of little value if not subsequently reinserted in the interactional context of discourse. In § 4, new avenues related to text message usage/analysis are explored.

20. The Japanese concordancing computer program *Casual Conc* (version 1.9.3), devised by Yasu Imao, Osaka University, was used for this purpose.

21. Other variations also appear: “je t aime” (610 tokens), “je taime” (104 tokens), “jtm” (73 tokens), “JTM” (2 tokens), “je tm” (7 tokens), “jtme” (1 token), “i love you” (3 tokens). Smileys such as “<3” (2,144 tokens), simulating a heart, also appear.

4. From ‘isolated’ to ‘conversational’ text messages

The current *sud4science LR* corpus includes solely ‘one-way’, isolated text messages. It is highly important to have this authentic data available, but it is nevertheless frustrating for linguists (also for computational linguists and specialists of other fields, including psychologists), who are interested in focussing on the ‘two-way’ or ‘conversational’ aspects. If bidirectional text-messages were to be available, sequences of exchanges between parties (such as those appearing in Figure 5 below) could then be analysed.

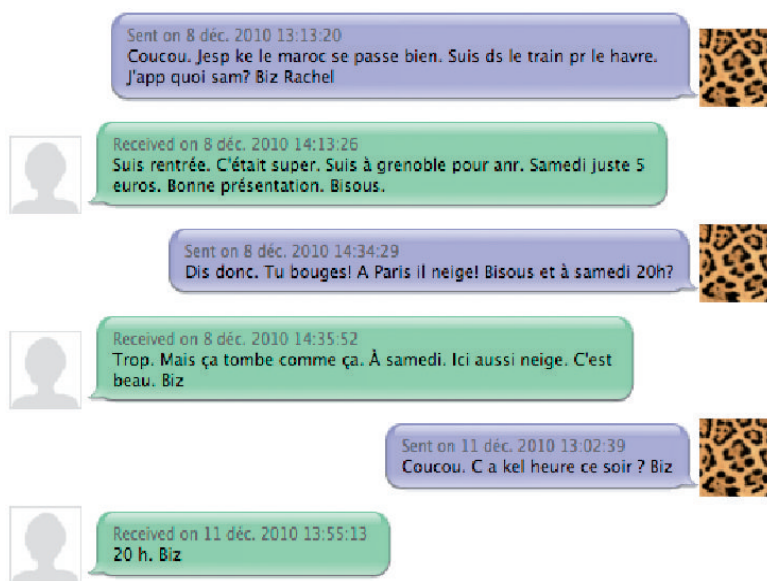


Figure 5. An SMS “conversational” exchange in French²²

A further SMS data collection is programmed for 2014–2015, which will allow comparison of these two types of data. In the following paragraphs, we address several issues of interest for this up-and-coming research. To our knowledge,

22. A. Hi. I hope Morocco is going well. Am in the train for Le Havre. What shall I bring on Saturday? Kisses. Rachel.

B. Am back. It was great. Am in Grenoble for a research project. Saturday, just 5 euros. Have a good talk. Kisses.

A. You're really on the move! It's snowing in Paris! Kisses and see you on Saturday. 8pm?

B. Too much. But that's the way it is. See you on Saturday. It's snowing here too. It's beautiful. Kisses.

A. Hi; What time this evening? Kisses.

B. 8 pm. Kisses.

conversational SMSes have never been collected in a scientific research project. It is of course very different to study language as such, out of context, compared to authentic interactions: in the latter, *openings* and *closures* can be analysed, as well as the way in which *themes* evolve, change, are relaunched; *discursive words* can also be studied along with specific marks of *expressivity* (C. Moïse & C. Schultz-Romain 2010). C. Fairon *et al.* (2006) indicate in the Belgian 2004 corpus radical change related to *conversational rules*: “flurries” of questions and answers, for example (see § 4.1., § 4.3. below).

4.1 Grammar and interaction

Grammar and interaction is at the heart of interactional analysis and it is both original and interesting to analyse conversations via SMS compared with face-to-face ones.

One of the aspects of interest concerning grammar and interaction is in relation with question/answer flurries and the way in which they appear to be entwined or out of sequence, as illustrated in the colour-coded sequence appearing in Figure 6 (R. Panckhurst & C. Moïse 2011, *cf.* note 23) — texter F (No 105) asks questions and gives information; texter B (No 106) decides not to answer one of the questions, but does react to private information and includes another topic opening.

105. F. Coucou L :) cmt tu vas ? quoi de beau ? Cmt se passe les cours ? Faudra se faire une soirée ! F est la ? Au fait scoop j'ai en coupe lol te rakterai ! Dis moi t'aurai le tel a C ? G a pas sn tel et elle connaît pas sn nim ! Biyoux

106. B. Salut ça va bien. Non je suis seule. Je croyais que vous m'avez oubliée !!! Cool je vois qu'on a des trucs a se raconter !!!! Non j'ai pas le num a C ! Dsl bonne journée, a bientôt ! Bizz

107. F. Lol nan tkt on ta pas oublié c juste quentre les cours et ma nouvelle relation j'ai pas trop trep de temps lol ^^ et ui c clair on a plein dtruc a se dire lol :) demain j'ai pas cours on peut se voir si tu veux :) [...]

Figure 6. Colour codes indicate the discursive question/answer “flurries”²³

23. 105. F. Hi L :) how are you? whaddya up to? how're lectures going? we'll have to get together one evening! Is F there? Btw scoop I'm in a relationship lol 'll tell you about it! tell me do you have C's number? G doesn't have her phone and she doesn't have his no! Kisses. 106. B. Hi I'm well. No I'm by myself. I thought you'd forgotten about me!!! Cool I see we've got things to tell each other!!!! No I don't have C's no! Sorry have a nice day, see you soon! Kisses 107. F. Lol no don't worry we haven't forgotten you it's just that with classes and my new relationship I don't have much time lol ^^ and yes for sure we've got lots of things to catch up on lol :) tomorrow I don't have lectures and we can get together if you like :) [...]

4.2 Context and conversational rituals

The *context*, whether temporal (“Si tu veux je viens te chercher à 6h”, *If you like I’ll come and get you at 6*) or spatial (“Je suis ds le train, todo bene.” *I’m in the train, all is well*), which is fundamental in order to fully understand usage and situate the exchange (M. Marcoccia 2011), is also easier to identify with preceding and following text messages (“suis à Grenoble”, *am in Grenoble*). When texting one another, it is typical for each party to provide information about his/her whereabouts.

Smartphones allow text message exchanges to visually flow like a discussion thread. Therefore, texters may not necessarily consider their message to be a *new* contact as such, but rather a continuation of previous exchanges. Politeness exchanges are not always necessary, and openings/closures (or lack of) show this. In a recent exploratory study of conversational text messages (R. Panckhurst & C. Moïse 2011), *closures* (“Gracias la madre, bonne semaine” *Gracias la madre, have a good week*) were more common than *openings*. This is also the case in the isolated *sud4science LR* corpus: of those lexical items studied (openings: “Bonjour”, “bjr”, “hey”, “coucou”, “cc”, “ça va”...; closures: “à demain”, “à toute”, “bisous”, “bonne soirée”...), 75% were closures compared to only 25% of openings. Comparing this aspect, among others, with a large conversational SMS corpus would be interesting, in order to check and see differences between the onset of a totally new exchange (as compared to subsequent openings within a wider exchange) and main closures as opposed to internal ones.

4.3 Events and emotions

Potential avenues the authors have envisaged include discursive types (from informational to appreciative), modes of expression, pragmatic and discursive processes which are typical of this type of communication (for instance, how does one express injunction? Are modalisations used as frequently as in daily face-to-face conversation?)

Text messages are often sent within a private sphere between texters who know each other. Events and shared emotions, which help build interactions, frequently appear in SMS writing, and are therefore interesting aspects to study.

Themes/topics are either unique (e.g. questions followed by answers) or in “flurries”: “Hey ca va? Alors cette rentree? Va falloir se trouver un p’tit creneau pour se voir! Dis moi, est-ce que tu sais quand commence les cours de langue non spécialiste? Bisoux” (*Hi, how’s it going? Back to university? We’ll need to find time to catch up! Btw, do you know when non-specialist language classes start? Kisses.*).

Emotions are often indicated via non-verbal semiological aspects (smileys) and character, punctuation, phoneme repetition (“suuuuppeeerrr”, “!!!!”, “ouiiii”)

simulating intonation, i.e. paraverbal information; interjections (“Hey”, “ben”), discourse words (“lol”), onomatopoeia (“grrr”), etc. also indicate these.

Conclusion

Overall, the organisation of the *sud4science LR* data collection project required a relatively significant time span, over a year in this case. Technical, legal, media, experimental and even interpretative aspects were all covered. Owing to this work, and the *anonymisation* phase, the researchers will now be able to carry out the *transcoding* phase and in-depth analysis of the text message data, within the framework of the *DGLFLF* project (cf. note 1) and also reflect upon collecting novel data for conversational exchanges in 2014–2015, which is the next step.

Conversational text messages can be collected if two people accept to donate their exchanges together. Anonymity is not preserved in the sense that in order to do so, subjects need to donate via a piece of software (Phoneview)²⁴ in the presence of the researchers conducting the data collection. The software allows the researchers to extract individual conversations without the telephones having to be entirely synchronised with the computer. Conversational corpora of this kind would be novel — as far as the authors know — and this specific data collection would provide interesting new paths for multidisciplinary analysis.

This article has covered a large quantity of ‘behind-the-scenes’ information, which may initially appear, in some cases, superfluous. However, the authors, and indeed the whole *sud4science LR* group, consider these ‘organisational’ and practical ‘structural’ aspects to be of significant importance in building the foundations of any scientifically ambitious project. Subsequent articles will deal with detailed analyses of the *sud4science LR* corpus, which, even though it currently contains solely “one-way”, “isolated” messages, is nevertheless a rich contribution to the global SMS authentic database, and will provide sources for authentic data analysis covering several discipline areas for years to come.

In the current era, people are writing (especially via mediated electronic discourse methods, including text messaging) more than ever before. Outstanding creativity and playful methods are being used every day by the young and the not-so-young, who are constantly inventing and reinventing scriptural practices which imply day-to-day mutations. It is a privilege to be able to analyse authentic SMS data and thus witness language (r)evolution. The authors agree with D. Crystal (2008, p. 175): “Some people dislike texting. Some are bemused by it. Some love it. I am fascinated by it, for it is the latest manifestation of the human ability to be

24. <http://www.ecamm.com/mac/phoneview/>

linguistically creative and to adapt language to suit the demands of diverse settings. In texting we are seeing, in a small way, language in evolution.”

Acknowledgements

The *sud4science LR* project is part of a vast international SMS data collection project, entitled *sms4science* (<http://www.sms4science.org>), and started at the CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium) in 2004. Our work is supported by the MSH-M (Maison des Sciences de l’Homme de Montpellier) and the DGLFLF (Délégation générale à la langue française et aux langues de France).

References

- Accorsi, Pierre; Namrata Patel; Cédric Lopez; Rachel Panckhurst; Mathieu Roch. 2012. Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques, *Lingvisticae Investigationes*, Special Issue: SMS Communication: A Linguistic Approach, John Benjamins, 35: 2, p.
- Anis, Jacques. 1999. *Internet, communication et langue française*. Paris: Hermès.
- Anis, Jacques; Michel de Fornel; Béatrice Fraenkel. 2004 (organisers). *La communication électronique: Approches linguistiques et anthropologiques*, Colloque international, EHESS, Paris, 5–6 February 2004.
- Antoniadis, Georges; Gaëlle Chabert; Virginie Zampa. 2011. *Alpes4science: Constitution d’un corpus de SMS réels en France métropolitaine*, talk, 79th Acfas colloquium, Sherbrooke, 9–10 May 2011.
- Bernicot, Josie; Olga Volckaert-Legrier; Antonine Goumi; Alain Bert-Erboul. 2012. Forms and functions of SMS messages: A study of variations in a corpus written by adolescents. *Journal of Pragmatics* 44: 1701–1715.
- Cougnon, Louise-Amélie; Thomas François. 2011. Étudier l’écrit SMS. Un objectif du projet *sms4science*, in Adrian Stähli / Christa Dürscheid / Marie-José Béguelin (eds) (2011), *La communication par SMS en Suisse. Usages et variétés linguistiques*. Linguistik Online (Themenheft).
- Cougnon, Louise-Amélie; Gudrun Ledegen. 2010. “c’est écrire comme je parle”. Une étude comparative de variétés de français dans l’écrit sms”. *Les voix des Français. Modern French Identities* 2(94): 39–57.
- Cougnon, Louise-Amélie. 2012. *L’écrit sms. Variations lexicale et syntaxique en francophonie*. PhD, Université Catholique de Louvain, September, 2012.
- Crystal, David. 2008. *Txting the Gr8 Db8*. Oxford: Oxford University Press.
- Détie, Catherine; Bertrand Verine. 2012. Quand l’insulte se fait mot doux: la violence verbale dans les SMS, Colloquium *Dimensions du dialogisme 3: Du malentendu à la violence verbale*, Helsinki, Finland, 15–17 August, 2012.
- Develotte, Christine; Richard Kern; Marie-Noëlle Lamy. (coord.). 2011. *Décrire la conversation en ligne*. Lyon: ENS Editions.

- Dürscheid, Christa; Elisabeth Stark. 2011. Sms4science: An International Corpus-based Texting Project and the Specific Challenges for Multilingual Switzerland. In *Digital Discourse. Language in the New Media*, C. Thurlow, K. Mroczek (eds.), 299–320. Oxford: Oxford University Press.
- Fabre, Ludivine; Manon Ravel. 2011. *L'étude des conversations SMS français / français et l'utilisation des Smartphones*, Master's assignment, Université Paul-Valéry Montpellier 3.
- Fairon, Cédric; Jean René Klein. 2009. La communication par SMS. *La lettre des académies* 13: 6–8.
- Fairon, Cédric; Jean René Klein. 2010. Les écritures et graphies inventives des SMS face aux graphies normées. *Le français d'aujourd'hui. Graphies : signes, gestes, supports*, 170: 113–122.
- Fairon, Cédric; Jean René Klein; Sébastien Paumier. 2006a. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*. Louvain-la-Neuve: Presses universitaires de Louvain, Manuel+CD-Rom. <http://www.smspourlascience.be/>
- Fairon, Cédric; Jean René Klein. Sébastien Paumier. 2006b. *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête "Faites don de vos SMS à la science"*. Louvain-la-Neuve: Presses universitaires de Louvain. <http://www.smspourlascience.be/>
- Gadet, Françoise. 2008. L'oral et l'écrit dans les changements technologiques et idéologiques, in E. Galazzi et C. Molinari (dir.), *Les français en émergence*. Bern: Peter Lang.
- Guimier de Neef, Émilie; Jean Véronis. (organisers). 2004. Journée d'Etude. Association pour le Traitement Automatique des Langues (ATALA). "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)", 5 juin 2004, ENST, Paris, Online publication. <http://www.up.univ-mrs.fr/~veronis/je-nfce/theme.html> [consulted on 7/4/11].
- Guimier de Neef, Émilie; Arnaud Debeurme; Jungyeul Park. 2007. TiLT correcteur de SMS: évaluation et bilan qualitatif. *Proceedings, 14th Conference on Natural Language Processing (TALN)*, 123–132.
- Herring, Susan C. (ed.). 1996. *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*, Amsterdam: John Benjamins.
- Langlais, Philippe; Patrick Drouin; Amélie Paulus; Eugénie Rompré Brodeur; Florent Cottin. 2012. Text4Science: a Quebec French database of annotated short text messages. *Proceedings, LREC*, May, 1047–1054.
- Liénard, Fabien. 2005. Langage texto et langage contrôlé. Descriptions et problèmes. *Linguisticae Investigationes* 28(1): 49–60, Paris.
- Liénard, Fabien. 2008. Analyse linguistique et sociopragmatique de l'écriture électronique. In *La langue du cyberspace: de la diversité aux normes*, J. Gerbault (éd.), Le cas du SMS tchaté, 265–278, Paris: l'Harmattan.
- Liénard, Fabien. 2011. Le SMS à Mayotte: le cas d'un écrit ordinaire quotidien favorisant la pratique des langues, dans Actes du colloque international. In *Plurilinguisme, politique linguistique et éducation. Quels éclairages pour Mayotte?*, F. Laroussi, F. Lienard (eds.), 99–120, Mont-Saint-Aignan: PURH.
- Marcoccia, Michel. 2011. "T'es où maintenant ?", les espaces de la conversation visiophoniques en ligne. In *Décrire la conversation en ligne*, C. Develotte, R. Kern, M.-N. et Lamy (eds.), 95–116, Lyon: ENS Éditions.
- Marty, Noëlle. 2005. *Informatique et nouvelles pratiques d'écriture*. Paris: Nathan.
- Matalon, Bernard. 1996. *La construction de la science. De l'épistémologie à la sociologie de la connaissance scientifique*. Neuchâtel: Delachaux Niestlé.

- Moïse, Claudine; Christina Schultz-Romain. 2010. Violence verbale et listes de discussions: les argumentations polémiques. In *Du "terrain" à la relation : expériences de l'internet et questionnaires méthodologiques*, I. Pierozak (ed.), Cahiers de l'institut de linguistique de Louvain: numéro 36/2, 113–133.
- Ouillon, Conrad. 2008. *Mieux soutenir le développement de la compétence à écrire*, Rapport du Comité d'experts sur l'apprentissage de l'écriture, Janvier 2008, <http://www.mels.gouv.qc.ca/>, fichier intitulé *SoutenirDeveloppementCompetenceEcrire.pdf* (consulted on 11/11/10), 49 pages, Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport 07–01107, ISBN 978-2-550-52096-2 (Print version) ISBN 978-2-550-52097-9 (PDF), Bibliothèque et archives nationales du Québec.
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In *Polyphonies*, pour Michelle Lanvin, T. Arnavielle (coord.), 33–52, Université Paul-Valéry Montpellier 3. <http://hal.archives-ouvertes.fr/hal-00443014>.
- Panckhurst, Rachel. 2010a. Texting in three European languages: does the linguistic typology differ?. *Proceedings i-Mean 2009 Issues in Meaning in Interaction*, University of the West of England, Bristol, <http://www.uwe.ac.uk/hlls/llas/events/0809/i-mean/index.shtml>, April, 2009. <http://hal.archives-ouvertes.fr/hal-00443016>
- Panckhurst, Rachel. 2010b. *SMS et eSMS : une autre forme de communication électronique médiée?*, invited conference, Psychologie du Développement et Processus de Socialisation (EA 1687) Laboratory, Université de Toulouse 2, 17 November. <http://hal.archives-ouvertes.fr/hal-00531592/fr/>
- Panckhurst, Rachel; Claudine Moïse. 2011. *SMS "conversationnels": caractéristiques interactionnelles et pragmatiques*, talk, 79th Acfas colloquium, Sherbrooke, 9–10 May 2011.
- Panckhurst, Rachel; Claudine Moïse. 2012. *sud4science Languedoc Roussillon, collecte de SMS isolés et conversationnels. Démarche et méthode scientifiques*, talk, VALS-ASLA colloquium, Lausanne, 1–3 February, 2012.
- Panckhurst, Rachel. 2012a. *sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée*, Presentation of the programme, Praxiling, 24 April 2012.
- Panckhurst, Rachel. 2012b *sud4science Languedoc-Roussillon: collecte et analyse de 90 000 SMS authentiques*, invited conference, La Fabrique-Mons (Belgium), 27 April 2012, <http://www.aprem-exploration1.net/programme.html>.
- Panckhurst, Rachel. 2012c. *Un nouveau corpus de SMS. La collecte sud4science LR.*, Oral online talk, Corpus Écrits — Nouvelles Formes de Communication, 31/5/12, slides: <https://groupes.renater.fr/wiki/corpus-ecrits-nouvcom/public/reunion-31mai/prog>.
- Panckhurst, Rachel; Catherine Détrie; Cédric Lopez; Claudine Moïse; Mathieu Roche; Bertrand Verine. 2014. A corpus of 88,683 authentic text messages in French. Programme 1): *sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée*, <http://www.msh-m.fr/programmes-2012/sud4science-lr/>; Data collection, online enrolment & consent, sociolinguistic questionnaire: <http://www.sud4science.org>; video presentations of scientific seminars: <http://www.msh-m.tv/spip.php?rubrique138>; Programme 2): *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS*, (Délégation générale à la langue française et aux langues de France, DGLFLF).
- Singly (de), François. 1992. *L'enquête et ses méthodes: le questionnaire*, Paris, Nathan: 122–123.

- Véronis, Jean; Émilie Guimier de Neef. 2006. Le traitement des nouvelles formes de communication écrite. In *Compréhension automatique des langues et interaction*, G. Sabah (ed.), 227–248, Paris: Hermès Science.
- Volckaert-Legrier, Olga; Josie Bernicot; Alain Bert-Erboul. 2009. Electronic mail, a new written-language register: A study with French-speaking adolescents. *British Journal of Developmental Psychology* 27: 163–181.

Projects/programmes:

- CENTAL: sms4science: <http://www.sms4science.org> [The original Belgian SMS data collection which took place in 2004. This website links up to the other *sms4science* websites: Reunion Island, Switzerland, Quebec, Rhône-Alps, Montpellier, see note 7].
- MSH-M: *sud4science Languedoc-Roussillon. Mutation des pratiques scripturales en communication électronique médiée*, January 2011–December 2012: (<http://www.msh-m.fr/programmes-2012/sud4science-lr/>); Data collection, online enrolment & consent, sociolinguistic questionnaire : <http://www.sud4science.org>; video presentations of scientific seminars: <http://www.msh-m.tv/spip.php?rubrique318>. [This programme focuses on SMS data collection, anonymisation and seminars and workshops conducted in Montpellier with international experts in the field].
- DGLFLF: *Pratiques contemporaines de la textualité numérique : observation, description et analyse d'un grand corpus de SMS*, (Délégation générale à la langue française et aux langues de France, DGLFLF), 1/7/12 to 30/6/13. [This project enables the researchers to move from the *anonymisation* stage to the *transcoding* stage of the sud4science LR corpus before conducting analysis of the text message data and the related sociolinguistic questionnaire.]

Summary

French text messages. From SMS data collection to preliminary analysis.

Over a three-month period (spanning 15 September to 15 December 2011), over 90,000 authentic text messages in French were collected by a group of academics in the Languedoc-Roussillon region of France. This paper retraces the organisation of the data collection, the elaboration of the sociolinguistic questionnaire that donors were invited to fill out, text message data processing procedures and preliminary results. A shift from individual “isolated” text messages to “conversational” SMS exchanges is then studied, in preparation for a new SMS conversational data collection which is due to take place in the near future. This whole process is important for understanding in-depth interactional practices within contemporary digital textuality and should provide insight for pluri-disciplinary approaches.

Appendix: Questionnaire

sud4science questionnaire Progression : 0 % Partie Introduction	
Vous pouvez choisir entre : donner votre consentement seulement (obligatoire) donner votre consentement (obligatoire) et remplir le questionnaire (optionnel)	
- Il n'est pas obligatoire de répondre au questionnaire, mais nous souhaitons que vous le fassiez, car les réponses aux questions sont très importantes pour notre projet. - Les informations que vous nous confiez seront utilisées de manière totalement anonyme. Nous nous y engageons, comme en témoigne le formulaire de consentement proposé ci-après. Il sera donc impossible à quiconque de faire le lien entre vous et les données entrées ci-dessous. Dans ces conditions, vous pouvez répondre au questionnaire en toute sérénité. Par ailleurs, après avoir rempli le questionnaire, si vous nous envoyez au moins 5 SMS, vous aurez le droit de participer au tirage au sort hebdomadaire afin de gagner des cadeaux. - Attention ! Pour participer au tirage au sort hebdomadaire, il faut obligatoirement remplir le questionnaire.	
Valider et poursuivre	
Âge Quel est votre âge ?	20 ans
Numéro de portable Veuillez indiquer votre numéro de téléphone portable (France métropolitaine) sans code international, ni espaces. Par exemple : 0601020304 Indiquez à nouveau votre numéro de téléphone portable	
Sexe :	féminin masculin commentaire
Êtes-vous en situation de handicap ? Si oui, veuillez préciser :	oui non sourd déficient auditif aveugle déficient visuel handicapé moteur autre :
Langue(s) pratiquée(s) Dans quelle(s) langue(s) vous sentez-vous le plus à l'aise ? (Par exemple : français, espagnol, créole, arabe, berbère, wolof, LSF...)	choisir une langue choisir une langue choisir une langue choisir une langue
Langue(s) pratiquée(s) Quelle(s) langue(s) employez-vous le plus souvent ? Les réponses peuvent être différentes par rapport à la question précédente.	choisir une langue choisir une langue choisir une langue

Langue(s) maternelle(s) Quelle(s) est/sont votre/vos langue(s) maternelle(s) ? Les réponses peuvent être différentes par rapport aux deux questions précédentes		choisir une langue choisir une langue choisir une langue choisir une langue
Niveau d'étude actuel :		primaire secondaire : choisir un niveau supérieur : choisir un niveau
Profession Quel est votre métier ou votre secteur d'activité ? Vous pouvez cocher plusieurs réponses.		☐ administration ☐ agriculture-viticulture ☐ artisanat ☐ cadre privé / public ☐ commerce ☐ élève / étudiant ☐ employé / ouvrier enseignement : précisez ☐ métiers de la santé ☐ métiers du spectacle ☐ profession libérale ☐ retraité ☐ sans activité professionnelle autre : précisez
Code postal Quel est le code postal de votre domicile ?		exemple : 34000
Valider et poursuivre		
← Revenir à la page précédente		
Type de téléphone Quel type de téléphone possédez-vous ?		☐ téléphone « intelligent » tactile (smartphone de type iPhone...) ☐ téléphone « intelligent » avec un clavier (smartphone de type Blackberry...) ☐ téléphone « classique » autre : précisez
Avez-vous un abonnement avec SMS illimités ?		☐ oui ☐ non
Depuis combien de temps utilisez-vous les SMS ?		choisir la durée
Combien de SMS envoyez-vous, en moyenne, par semaine ?		choisir le nombre
Combien de SMS recevez-vous, en moyenne, par semaine ?		choisir le nombre
Utilisez-vous la correction automatique (T9, écriture intuitive) pour la rédaction de vos SMS ?		☐ oui ☐ non

Classez les destinataires par ordre de fréquence, de 1 à 6.

Utilisez « 6 » pour le plus fréquent, « 5 » pour le suivant et ainsi de suite. Notez « 0 » pour les éventuelles catégories qui ne vous concernent pas.

0 1 2 3 4 5 6
amis
le - le +

0 1 2 3 4 5 6
collègues

0 1 2 3 4 5 6
famille

0 1 2 3 4 5 6
petit(e) ami(e), compagnon / compagne

0 1 2 3 4 5 6
concours / jeux

0 1 2 3 4 5 6
autre, à préciser

Rédaction de SMS

En général, quand vous rédigez vos SMS, utilisez-vous l'orthographe « standard » ou l'« écriture SMS » ?

l'orthographe standard uniquement

l'écriture SMS uniquement

le mélange de l'orthographe standard et de l'écriture SMS

soit l'orthographe standard, soit l'écriture SMS selon le destinataire

l'orthographe standard pour les SMS formels (ex. : travail, professeur, etc.), et l'écriture SMS pour les SMS informels (ex. : amis, famille, proches, etc.)

Utilisez-vous d'autres systèmes de communication sur Internet ?

chat :

forum :

courriel électronique :

réseaux sociaux :

autre, à préciser :

Valider et poursuivre

← Revenir à la page précédente

Pourquoi écrivez-vous des SMS (au lieu de téléphoner, par exemple) ?

parce que c'est moins cher ou inclus dans le forfait
pour ne pas déranger
pour aller plus vite
parce que je n'aime pas téléphoner
autre :

Si vous rédigez en écriture SMS, pourquoi le faites-vous ?

parce que c'est plus rapide
parce que j'aime jouer avec la langue
parce que ça crée une complicité avec les amis
autre :

Opinions sur les SMS

D'après-vous, quels sont les avantages / inconvénients de l'écriture SMS ?

Avenir des SMS

Pensez-vous qu'à l'avenir les SMS seront encore utilisés ?

Merci d'expliquer votre réponse, en nous donnant, si possible, des exemples.

Changements et mélanges des langues

Quand vous rédigez vos SMS, mélangez-vous les langues ?

oui non

Si oui, lesquelles ?

choisir une langue

choisir une langue

choisir une langue

choisir une langue

Changez-vous de langue en fonction des destinataires ?

oui non

Si oui, précisez les langues utilisées :

choisir une langue

choisir une langue

choisir une langue

choisir une langue

Habitudes de lecture / écriture

Comprenez-vous facilement les SMS que vous recevez, s'il y a utilisation de l'écriture SMS ?

oui, sans problème

oui, mais je dois parfois « décoder »

non, c'est illisible

Vous vous souvenez plus facilement du contenu d'un SMS...

lorsque l'écriture SMS est utilisée

lorsque le français « standardisé » est utilisé

pas d'opinion

Combien d'heures par jour consacrez-vous à la lecture quelle qu'elle soit (en dehors de la lecture de SMS et de courriels) dans le cadre scolaire / professionnel ou de vos loisirs ?

Indiquez le nombre d'heures ci-contre, par exemple : 2, si vous lisez 2 heures par jour.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

0 h / jour

De quel type de lecture s'agit-il ?

des articles de journaux

des BD

des pages de sites Web

des livres

autre : précisez

Lors de vos activités professionnelles / scolaires et de vos loisirs, consacrez-vous du temps à écrire sur papier, de quelque manière que ce soit ?

oui non

Si oui, sous quelle forme ?

liste de courses

courrier

journal intime

petits mots

cours ou notes professionnelles

autre : précisez

Si oui, à quelle fréquence ?

choisir la fréquence

Valider et poursuivre

← Revenir à la page précédente

Je confirme que j'ai écrit moi-même les SMS envoyés. Je consens à ce que les SMS envoyés soient stockés dans le corpus SMS et à ce qu'ils figurent dans des publications sous forme anonymisée en tant que partie du corpus. Je consens à ce que les réponses au présent questionnaire soient stockées de manière anonymisée au sein du corpus de SMS et à ce qu'elles figurent dans des publications sous forme anonymisée.

Je refuse de participer au tirage au sort hebdomadaire.

Terminer

A sociolinguistic analysis of transnational SMS practices

Non-elite multilingualism, grassroots literacy
and social agency among migrant populations
in Barcelona*

Maria Sabaté i Dalmau
Universitat de Lleida (UdL)

Introduction: Towards a sociolinguistics of transnational migrants' SMS talk

In current network societies, the mobility of people and of all sorts of resources around the world is largely orchestrated by ICTs, information and communication technologies (A. Appadurai 1996; M. Castells 2000 [1996]; M. Castells *et al.* 2007a; J.E. Katz 2008; M. Sheller and J. Urry 2006). These have become particularly crucial for the organisation of the life trajectories and work prospects of transnational migrant populations,¹ who today are at the forefront of technology adoption (A. Alonso and P.J. Oiarzabal 2010; J. Blommaert 2012; H.A. Horst and D. Miller 2006; A. Ros *et al.* 2007; S. Vertovec 2006). Along with the Internet, the mobile phone has become one of the most popular and increasingly cheapest tools for

* This research was funded by research group grant FFI2011–26964 (MINECO). I'm thankful to the C.I.E.N. research group for their support. Any shortcomings are, of course, mine.

1. I employ the term *transnational migrant* for two reasons. Firstly, mobile citizens' trajectories today are highly complex and imply living in various nation-states simultaneously, or paying long and short visits to other places (C. Solé *et al.* 2007; C. Solé *et al.* 2008; S. Vertovec 2001). The term *migrant* better gathers the idea of groups of people continuously re-defining their mobility. Secondly, many informants did not want to be called 'immigrants', and some found the term pejorative and exclusionary. With the use of the term *migrant* I try to dissociate myself from a more traditionally ethnocentric approach to demographic movements. Following Vertovec (2009: i), I use the term *transnationalism* to make reference to "the multiple ties and interactions linking people or institutions across the borders of nation-states [...] within the study of globalization" — those which simultaneously have an effect on *both* tiny localities and big nation-states; on the global and the local.

establishing quick communication worldwide (N.S. Baron 2008; G. Goggin 2006; R. Ling and P.E. Pedersen 2006; H. Rheingold 2002), to the point that, over the last few decades, it has evolved into the new *social glue* (S. Vertovec 2009: 54) or *communication staple* (A.N. Panagakos and H.A. Horst 2006: 112) of transnational family units (J. Androutsopoulos 2006a, 2006b, 2007; B. Danet and S.C. Herring 2007; E.W.S. Lam and E. Rosario-Ramos 2009; E.W.S. Lam and D. Warriner 2012).

Despite the fact that the global ICT system has traditionally been geared towards a *technoliterate*² elite and has largely ignored the communicative needs of those mobile phone users who do not have access to the languages, alphabets and semiotic and alphanumeric signs of the Northern Hemisphere (see D. Cameron 2000; J. Chipchase 2008; G. Kress and T. van Leeuwen 2006 [1996]; M. Sabaté i Dalmau 2012a), migrants have massively appropriated the mobile phone's text messaging option (henceforth SMS) for accessing crucial transnational *social capital*; that is, for attaining more material and symbolic resources and, thus, increased networking power (I follow P. Bourdieu's notion of capital and resource here; see P. Bourdieu 1977, 1990, 1991). Thus, this particular type of ICT-mediated communication has become a very convenient tool for the successful articulation and protection of migrants' transnational trajectories in network societies like Catalonia, where the present study took place.

Catalonia is an autonomous Catalan-Spanish bilingual region located in the north-eastern part of Spain (see Generalitat de Catalunya 2006) where, as in most places in Europe, documented migrants, who accounted for 15.7% of a total population of 7,535,251 in 2011 (Idescat 2011), showed a remarkably higher connectivity rate when compared to Spanish nationals. At the time when this study was conducted, 96.2% of transnational migrants used the mobile phone as their primary means of communication (FOBSIC 2010), particularly for SMS communication (M. Castells *et al.* 2007b: 53). Besides, those aged between 15 and 29 actually initiated *twice* as many calls, SMS and Internet connections as non-foreigners, spending 40% more money on communication technology than Spaniards (J. Robledo 2008), and demonstrating that they had definitely gone mobile.

The fact that most migrant populations now massively rely on SMS communication for the organisation of everyday life is particularly interesting for a branch of sociolinguistics that today is broadly known as the sociolinguistics of globalisation (see J. Blommaert 2003, 2010; N. Coupland 2003, 2010; N. Fairclough 2006; M. Heller 2010), for two main reasons. Firstly, the investigation of transnational SMS talk can provide a new lens on the role of language and literacy in the social

2. Technoliteracy is defined as the command of a complex set of visual, auditory, iconic, digital, informational, and technological and computer multiliteracies required for ICT-mediated communication (M. Area Moreira *et al.* 2008; R. Kahn and D. Kellner 2005).

structuration (A. Giddens 1984) of highly mobile transnational migrant populations. This is especially so for the migrants' wide range of heteroglossic written multilingual practices, as well as for their technology-related grassroots or vernacular literacies and increased iconographic repertoires — here SMS language is understood both as real situated linguistic practices which are messy, multi-valued and ever-changing and as empirically observable ideologies and discourses which are an indicator in social contestation, reproduction and change (following the frameworks posited by A. Duranti 1997, M. Heller 2007, and B.B. Schieffelin *et al.* 1998, among others). Secondly, the investigation of transnational SMS can also become a window into the technoliteracy practices that mobile phone users from very different migrant backgrounds mobilise as a way to establish, develop and maintain social relationships and identity affiliations and to re-negotiate their belonging to the host societies (see, also, J. Androutsopoulos 2012; E.W.S. Lam 2009; L. Martín-Rojo and C. Molina 2012; M. Sabaté i Dalmau 2012b, 2012c).

However, little is known as yet about the transnational text messaging practices conducted by highly multilingual migrants, since the growing body of research on twenty-first century SMS communication and on ICT-mediated talk “has largely neglected migrant and diasporic contexts” (J. Androutsopoulos 2006b:522; see, also, M. Jacquemet 2010). This is because the more traditional line of investigation in SMS language practices first focused on highly literate and educated middle-class teenagers and young adults born and socialised in Western societies; notably in the United States, the United Kingdom, Norway, Sweden, Finland and Japan — the cradles of the texting culture (R.E. Grinter and M. Eldridge 2003; J.E. Katz and M. Aakhus 2002). These first studies basically focused on form and approached SMS language from a largely descriptive point of view, codifying or “translating” the sloppy, telegraphic textese varieties full of acronyms, smiley codes and sophisticated abbreviations (see, by way of example, the studies by N.S. Baron 2008; E. Borochofsky-Bar-Aba and Y. Kedmi 2010; D. Crystal 2008; M. Torres i Vilatarsana and L. Payrató 2003). This research tradition saw linguistic codes as quantifiable, bounded, a-social units, and normally took monolingualism and the established standard norms of given codes (grammatical, orthographic, stylistic or otherwise) as the default legitimate point of reference for the analysis of SMS language, giving space, in its beginnings, basically to English-language investigations (B. Danet and S.C. Herring 2007; C. Thurlow and K. Mroczek 2011a, 2011b). More recent studies moved away from these more traditional approaches and focused on the vast array of indexicalities or social meanings of SMS language by linking text messaging with other situated communicative, linguistic and literacy practices (J. Androutsopoulos 2006a, 2006b; E.W.S. Lam 2009; E.W.S. Lam and E. Rosario-Ramos 2009; E.W.S. Lam and D. Warriner 2012), and by establishing newer community- or group-centred approaches that opened up a venue for the

more comprehensive investigation of bilingual and multilingual text messaging practices from a non-normative sociolinguistic and discourse analytic perspective (see the work by J. Androutsopoulos 2012; B. Danet and S.C. Herring 2007; A. Deumert and S.O. Masinyana 2008; R. Ling 2004, 2005; C. Tagg 2012; C. Thurlow and A. Brown 2003; and C. Thurlow and K. Mroczek 2011a, to mention but a few). These latter ethnographically-grounded perspectives and qualitative approaches to SMS talk have set the theoretical underpinnings of, and have become the point of departure for, this investigation, where I contribute to the studies of the societal phenomenon of SMS communication gathered in this book by working towards a very particular and largely unexplored arena: a critical sociolinguistic ethnographic analysis of transnational migrants' SMS talk, as presented in more detail in the following section.

1. Aims of this study, method and data

Following pioneering ethnographies and discourse analytic studies of mobile phone and Internet language usage investigating the locally situated communicative practices in both online and offline discursive spaces from an emic approach (J. Androutsopoulos 2012; H.A. Horst and D. Miller 2006; D. Slater and D. Miller 2000), in this paper I analyse the linguistic features, the social indexicalities and the particular uses of seven ordinary transnational text messages. These were type-written and sent by three migrant men of diverse ethnolinguistic, class and educational backgrounds, either to their non-foreign partners or to the researcher, at a time when they were establishing their transnational networks in a very marginal neighbourhood in the metropolitan area of the Catalan capital, Barcelona.

Taking these SMSes as my point of departure, in the analysis section I first describe some of the common linguistic characteristics of transnational text messages, which include, for instance, type-written neographical transformations like heterography or amalgamated spellings. I show that many migrants' SMS communicative practices consist of a nuanced use of multimodal flexibilised codes, which are circulated as "zero point codes" aimed at establishing a common ground for successful inter-group communication and for ICT-mediated cooperation across diverse migrant networks (J. Blommaert 2012; J. Blommaert and B. Rampton 2011). These migrants' written *linguae francae* are hybrid by their very nature and bear unique traces of non-elite multilingualism and grassroots or vernacular literacies, including many orality traits and anti-standard, inconsistent spelling. I argue that these are fully-fledged ecumenical "we-codes" (J.J. Gumperz 1982) whereby migrants gate-keep, self-distribute and gain access to key material and symbolic resources among themselves and in their own discursive spaces, outside

the sociolinguistic hierarchies and the linguistic *regimes* (P.V. Kroskrity 2000) established in the Catalan network society, which tends to foster a monoglossic unified floor in standard Spanish as the legitimate mainstream form of SMS communication (M. Sabaté i Dalmau 2012a, 2012c).

In the second part of the analysis section, I focus on the discourses and on the linguistic ideologies surrounding transnational migrants' SMS practices. I first describe the pejorative meta-comments of some of the SMS receivers. I then approach this censorious rhetoric as a set of illustrative examples of the ethnocentric beliefs surrounding migrants' technoliteracy capitals and of the conceptualisations of what shall count as legitimate "languageness" and as proper transnational SMS talk (frequently boosted by the negative stance on SMS language taken by the mass media; C. Thurlow 2006). I show that, more often than not, the migrants' unconventional and fluid type-written exchanges are seen as instances of a poor or "faulty" command of the established language and literacy norms employed by migrant mobile phone users who get categorised as "deviant" and "incompetent" (J.J. Gumperz 1986). Against this stereotyping regime of thought, I argue that the productive grassroots informalisation of standard norms in transnational SMSes is a very useful and empowering resource that provides migrant texters with the degree of *social agency* (A. Giddens 1984) which is necessary to make their own voice heard within and across different transnational networks. I explain that by mobilising this heteroglossic and hybrid SMS talk, migrant texters invest in multiple acts of identity and largely successfully develop different linguistic strategies for showing group belonging (or non-belonging). That is, I claim that transnational SMS language empowers migrants with a new linguistic means of resistance to articulate their marginal social position in their host societies, to re-negotiate their entitlement to city resources and, in fact, to institutionalise, via the ICTs, their more horizontal "nonofficial ways of speaking" (C. Thurlow and M. Mroczek 2011b: xxiii).

In the concluding section, I argue that the transnational text messages described in this article are just a very small window into the still largely unknown informal non-elite multilingual realms and grassroots literacy practices of the twenty-first century. I also emphasise that the migrants' SMS communicative practices are only one of the many ways in which migrants demonstrate that they can successfully transform, appropriate and autonomously insert themselves into the established contemporary communicative and literacy regimes of a Western-based network society, in a bottom-up manner and from the margins. I finish my reflections with a word of caution. I state that these transnational texting practices, despite their empowering potential, are far from liberating, since, as M. Jacquemet (2010:63) aptly points out, "they are still inserted into a global indexical order which assigns superior values to certain systems of communication" — in our case, into a shared monoglossic regime of mind in formal standard Spanish, to the

detriment of the migrants' allochthonous minority codes, which remain silenced both in the social media and in the telecommunications worlds.

The data presented in this paper are part of a *network ethnography* (Howard 2002; M. Sabaté i Dalmau 2012b) of a very unique transnational space of ICT communication, a *locutorio* or migrant-tailored cybercafé, conducted over two years of fieldwork (2007–2009). During such time, I investigated the emergence and consolidation of these alternative call shops which in Catalonia (and probably in other countries) have become migrant institutions of resistance against social exclusion and marginalisation, since today they offer resources which are essential for transnational survival (for instance, they sell mobile phone SIM cards to the undocumented, they offer cheaper international calls, they provide key knowledge and information about the host society in the migrants' languages, they distribute informal legal advice, job offers, rooms for rent, shelter when it rains, and so on; see M. Sabaté i Dalmau 2012c). The particular *locutorio* that I observed was located in a migrant-populated marginal neighbourhood on the outskirts of the metropolitan area of Barcelona, in a traditionally working-class county called Vallès Occidental, where I was living at the time of the fieldwork. There, the linguistic market was, and still is, secured in non-standard Spanish (K.A. Woolard 1989, 2003), a language which dominated and figured highly over Catalan, the co-official language of Catalonia, and over any other linguistic code.

During these two years, I was able to observe, *in situ*, the uses of landline telephony, Internet and SMS services by a very heterogeneous transnational network of 20 migrants aged between 27 and 52, born in Pakistan, Morocco, Romania and, to a lesser extent, Argentina, El Salvador, the Dominican Republic, Bolivia, Guinea, Equatorial Guinea, Cuba and Brazil. They were mostly unemployed or non-formally employed, and they all visited the cybercafé on a daily basis in order to establish or maintain crucial networks and, therefore, increased social capital.

Out of the 28 transnational SMSes sent by the members of the aforementioned network that I recorded during the fieldwork, here I only provide an in-depth qualitative and non-exhaustive analysis of seven illustrative text messages written by three of the most active migrant participants, reproduced *verbatim* with informed consent. These three informants are: Mohammed from Morocco, Naeem from Pakistan, and Shabbir from Kashmir.³ Their SMSes are all non-public and intimate, which better reflects real ICT-mediated language practices (S. Peuronen 2011:159). They respectively deal with an intimate declaration of love by Moha,

3. For confidentiality reasons, all names are fictitious and all places have been anonymised, as agreed with the Committee for Ethics in Animal and Human Research (CEEAH) at the Universitat Autònoma de Barcelona (file 725H) and with the cooperative of lawyers *Col·lectiu Ronda*.

a desperate cry against exploitation at work, three routine “take good care” and one thank-you message by Naeem, and, finally, a personal Christmas greeting by Shabbir.

I was able to follow the three informants in their workplaces, in the bar, around the neighbourhood and even in their households, on a daily basis for at least three months, and at least three times a week for the rest of the two years. Thus, I was able to complement my analysis of transnational SMS talk with the data that I also gathered on the SMS receiver’s reactions, comments and interpretations of the migrants’ text messages in the long term.⁴ In addition, I also draw on other locally-situated recorded communicative practices conducted by the informants in the same space (like hand-written notes), following the necessary “shift from ‘medium-related’ to more ethnographically grounded ‘user-related’ approaches” (C. Thurlow and K. Mroczek 2011b:xxi), and providing a broader (yet, of course, by no means complete or exhaustive) picture of some of the key sociolinguistic features of ICT-mediated non-elite multilingual and multimodal language in migration contexts, which I analyse in detail in the following section.

2. Migrants’ transnational SMS practices: Heterography, grassroots literacy and other transidiomatic practices

I start the analysis of the linguistic features of transnational SMS with a text message that informant Mohammed (“Moha”) sent to his partner Jenny on 29 July 2008. Moha was a documented blue-collar migrant born in Morocco, who worked in an electrical appliances factory in the metropolitan area of Barcelona and who could text in Spanish and in Romanised Arabic. Jenny, the SMS receiver, was a 35-year-old woman self-categorised as a Spanish-using “*gitana catalana*” (“Catalan gipsy”) who was born in town and who was informally employed as a cleaning person there. She kindly interpreted the SMS with me, which is reproduced *verbatim* in (1).

- (1) “*Ola carinyo estue mue continto para verte corazon estamos juntos y na braso*”
The transcription in standard peninsular Spanish (my version):
“*Hola cariño, estoy muy contento de verte, corazón. Estamos juntos. Un abrazo.*”
A possible translation into standard English (my version):
“*Hi love, I’m very happy to see you, sweetheart. We are together. A hug.*”

4. The researcher used Spanish, English and Catalan with the informants (only some chunks in Urdu, Punjabi and Arabic were employed over these two years).

Mapped upon Spanish, Moha's text message shows some transcodic marks and some traces of grassroots or vernacular literacy and of a "low" command of the conventional graphic representations of the standard variety of this particular language. For example, the SMS includes *heterography*, or "the deployment of graphic symbols that defy orthographic norms" (J. Blommaert 2008: 7), which can be seen in some of his spelling difficulties, as in his use of the adjective "*continto*" (instead of "*contento*"). Besides, he shows some traces of *xenoglossy* as well, understood as the linguistic ability to access and to appropriate a code without fully understanding it or without having fully learned it (M. Jacquemet 2005; J.N. Jørgensen *et al.* 2012). This occurs when he writes "*na brazo*" (arguably equivalent to "an arm"), possibly a morphological reinterpretation of "*abrazo*" ("hug"), because both the standard and the unconventional forms (that is, both "*un abrazo*" and "*na brazo*") are phonetically equivalent. Due to his multiple literacies, he also employs transcodic marks from Arabic phonology that may reflect "accent", as seen in the verb "*estue*" or in the adverb "*mue*". Thus, this SMS also contains a fair amount of orality, demonstrating that the text sender enjoys the current post-literate technological context which allows for meaning-making across oral and written modes (A. Pennycook 2007).

These features (heteroglossic and xenoglossic language, orality traits and transcodic marks) are all examples of the *transidiomaticity* that migrants routinely mobilise as a successful resource for ICT-mediated transnational navigation, networking and survival. According to M. Jacquemet (2005, 2010), these transidiomatic practices are the result of the multilingual talk put into motion by de- and re-territorialised peoples with differing linguistic capitals who interact via ICTs. In other words, they are "the communicative practices of transnational groups [...] using different languages and communicative codes simultaneously present in a range of communicative channels, both local and distant" (M. Jacquemet 2005: 264–265).

By going beyond the mere linguistic description and by linking the analysis of Moha's transidiomatic practices to the social indexicalities of his type-written communicative repertoires, we can see that Moha shows a unique command of the particular type of language employed for in-group membership and proximity, and for intimacy and love. With this hybrid talk, he manages to semiotically and aesthetically transmit his spoken words (that is, his intimate *voice*) to his partner Jenny, who in response has also started introducing transcodic intimate language games into the Spanish SMS that she sends him. For example, she explains that, now, when she texts him back, she consciously includes the expression "*mahibek*" in Romanised Arabic, which means "I love you", in a systematised way (as a farewell).

Moha uses orality traits through the absence of punctuation and orthographic conventions (like accents or commas), too, and through very common shortened

word forms employed for salutations, like “ola” (instead of “hola”), a popular informal greeting which, he claims, he chose on purpose. Thus, this type-written alternative language also includes iconicity and conscious and unconscious anti-normativity. I argue that this use of non-standardness is not only a mere type-written instrumental way to circulate information or to establish communication quickly, but also a powerful expressive resource which, according to A. Jaffe (2000: 498), has “the potential to challenge linguistic hierarchies” and to “make non-standard voices visible/audible in a medium that habitually does not recognize them”. That is, this fluid Spanish is part of the non-standard textese variety circulated among young migrant and non-migrant Spanish-using transnational networks gathered in the cybercafé of the neighbourhood that I investigated. More specifically, I read the SMS in (1) as Moha’s powerful “symbol of commitment” (J. Androutsopoulos 2007: 349) to his group and to his group’s identity practices, which involve a proud sentiment of being part of his partner’s Spanish-using core network of working-class youngsters who have become the most active and numerous in the call shops and the bars of the area.

All the 28 transnational SMSes that I recorded contain this type of hybrid, faulty ‘agrammatical’ language and this inextricable mixture of codes and overlapping communicative frames mapped upon a Spanish type-written linguistic landscape. The following text message (2), for instance, gathers the SMS practices of Naem, a 27-year-old cybercafé worker holding a temporary residency permit who came to Barcelona in 2007. Naem, whose linguistic capitals included a command of spoken and written Spanish, Urdu, Punjabi and English, was my window into the transnational network of migrants whom I investigated, for I spent most of my fieldwork time near the counter of his cybercafé. By the time he sent me the text message presented in (2) below (on September 30, 2008) he was undergoing a series of exploitative practices at work, which included extra shifts of 14 hours and three consecutive non-payments of his salary over the last three months, to mention but a few. He ended up leaving that workplace and found another job relatively soon, but during his last week in the cybercafé he desperately asked for his transnational network’s help, since in order to be finally granted permanent residency in Spain he needed to prove at least five years of work, which meant that he could not afford unemployment (like Moha, Naem also helped me interpret the SMS he sent me).

- (2) *“maria buenos noches yo otra base aqi pero no mas de un dia porfavor losiento muchesimo porfvor.ok by”*

The transcription in standard peninsular Spanish (my version):

“Maria, buenas noches. Yo estoy otra vez aquí, pero no más de un día por favor. Lo siento muchísimo. Por favor. Ok, adiós.”

A possible translation into standard English (my version):

"Maria, good night. I'm still again here but for no more than one day, please. I'm really sorry. Please. Ok, bye."

Naeem's SMS also shows transcodic marks where non-standard Spanish coexists with English (in this case), as seen with the non-standard farewell "*ok by*" ("bye"). The message bears traces of heterography and orthographic uncertainty as well, as found with his simultaneous use of and the hesitation between "*porfavor*" and "*porfvor*" showing non-conventional "polyvalent amalgamated spellings", including the suppression of vowels and the elimination of spaces (J. Anis 2007: 101, 104), which are found, too, in the speech act "*losiento*" ("I'm sorry", here written together as a one-word chunk) — these particular spellings were not a deliberate choice for making the SMS shorter though still (ortho)graphically intelligible, for Naeem kept insisting that he wanted to learn how to write "properly", and he frequently asked me to "correct" his written Spanish. Naeem also uses "truncations" (J. Anis 2007: 90); that is, words where letters are separated from each other, rather than joined together, as tends to happen in the Arabic script (D. Palfreyman and M. Al Khalil 2007: 49). More examples of truncations can also be found, for example, in his hesitation between "*vale*" and "*va le*" (which also mean "ok", in Spanish) and between "*he cho*" and "*hecho*" ("miss"), presented in the SMS reproduced *verbatim* in (3) with "*vale*"/"*hecho*" and in (4) with "*va le*"/"*he cho*" (respectively sent to me by Naeem on October 7 and November 28, 2008):

- (3) *"hola maria soy naeem como estas que todo bein te hecho de menios vale maria ok by cuidate"*

The transcription in standard peninsular Spanish (my version):

"Hola Maria. Soy Naeem. ¿Cómo estás? ¿Qué, todo bien? Te echo de menos. Vale Maria, vale adiós. Cuídate."

A possible translation into standard English (my version):

"Hi Maria. It's Naeem. How are you? Everything ok? I miss you. Ok Maria ok bye. Take care."

- (4) *"buenos dia maria yo tambien te he cho mucho de men ios cuidate mucho va le"*

The transcription in standard peninsular Spanish (my version):

"Buenos días Maria. Yo también te echo mucho de menos. Cuídate mucho, ¿vale?"

A possible translation into standard English (my version):

"Good morning, Maria. I also miss you lots. Take good care, ok?"

In the third place, Naeem's texting resources draw on some xenoglossic practices as well, particularly with morphological aspects that concern adjective-noun

agreement in Spanish, as his greetings illustrate in “*Buenos noches*” (instead of “*Buenas noches*”) in (2) and also in “*Buenos dia*” (instead of “*Buenos días*”) in the SMS presented in (4).

Naeem’s SMS practices also show that transnational migrants like him can mobilise inconsistent syntactic structures, unspecific or unconventionally inflected verbs and elliptical constructions to establish transnational communication successfully (see, also, Borochofsky-Bar-Aba and Kedmi 2010). Firstly, in (2), for instance, he constructs unfinished unconventional speech acts (greetings, requests and apologies, in this case) by making use of canonical — yet partial and very general — one- or two-word frequently used chunks of language which evoke Spanish orality (further evidence of his command and recurrent use of prefabricated introductory speech acts is found in the greeting “*hola maria soy naeem*” in (3)). These are clearly emphasised by means of repetition (instead of paraphrasing), particularly with the requests “*porfavor*”/“*porfvor*”, which in the end become clear and transparent supplicatory pleas. Secondly, he tends to omit existential verbs which require subject complement. More specifically, his elliptical constructions in my corpus basically consist of the omission of the copula verb “*to be*”, as seen, for instance, in “*yo otra base*” (literally, “I [omission] here again”) in (2), or in expressions like “*yo muy conteto*” (“I [omission] very happy”) in some of his other SMS (see (6) below). Thirdly, in the one case in which he did use the verb “*to be*”, he inflected it in an unconventional manner which arguably reflected an influence from the English language. This is seen in text message (5), where he typed “*ares*” instead of “*eres*” (for “*are*”), thus seemingly joining the second person singular inflectional systems of both English (“*are*”) and Spanish, which requires the inflectional ending “*-s*” (SMS sent to me on 30 July 2009):

- (5) “*Maria ares una buena chika cuidate mucho by*”

The transcription in standard peninsular Spanish (my version):

“*Maria, eres una buena chica. Cuídate mucho. Adiós*”

A possible translation into standard English (my version):

“*Maria. You are a good girl. Take good care. Bye.*”

Like Moha, Naeem shows a remarkable command of the textese variety and the street language of the neighbourhood, as seen for example with the use of emphatic informal spoken prefixes in Spanish, like “*-ísimo/a*”, (as in “*muchísimo*” in (2), equivalent to “*muchísimo*”, or “very”) and with the use of suffixes like “*super-*” (as in “*superbien*”, which means “super-well”), found in another SMS that he sent on 6 November 2008, reproduced *verbatim* in (6) — note that, with the use of “*Buenos noche*” and “*conteto*”, he resorts to xenoglossy again:

- (6) *“hola maria yo muy conteto que tu mi contesta te echo de menios tambien estoy superbien buenos noche!cuidate mucho vale”*

The transcription in standard peninsular Spanish (my version):

“Hola Maria. Estoy muy contento de que tú me contestes. Te echo de menos. Yo también estoy superbién. ¡Buenas noches! Cuídate mucho ¿vale?”

A possible translation into standard English (my version):

“Hi Maria. I’m very happy you replied. I miss you too. I’m super-well. Good night! Take good care, ok?”

Thus, Naeem’s SMSes show that he commands not only the informal street language but also some written formality strategies, denoting that he has mastered the use of a remarkably wide range of politeness markers and familiarity/formality registers. In his first SMS (2), for instance, he reveals a careful construction and organisation of the discourse. His introductory greeting, which is fully spelled out (*“maria buenos noches”*), is first followed by a request with the downtoner: “please” (*“no mas de un dia porfavor”*); then by the apologies for the immediacy of such a request (*“losiento muchesimo”*) as well as with the repetition of his plea (*“porfvor”*); and finally by a polite farewell, marked by means of a punctuation mark (a full stop), which also marks the switch of code from Spanish into English (*“ok by”*).

The SMS presented in (7) below allows for a more nuanced exploration of the presence of politeness markers and formality registers in transnational migrants’ SMS talk. It consists of an SMS (sent on 24 December 2008; also reproduced *verbatim*) provided by Shabbir, a 41-year-old construction worker who is competent in Urdu, Punjabi, English, Spanish, and some Catalan. Shabbir arrived in Barcelona in 2001, and he shared a flat with a Pakistani worker who had also left his extended family behind. The Christmas greeting that he texted me reads as follows (I read it again with him):

- (7) *“Hola.buens noche . srts maria mi Hermana felis novides.from m. Shabbir mughal”*

The transcription in standard peninsular Spanish (my version):

“Hola. Buenas noches, srta. Maria, mi hermana. Feliz Navidad. De Shabbir mughal.”

A possible translation into standard English (my version):

“Hello. Good night, Miss Maria, my sister. Merry Christmas. From M. Shabbir Mughal.”

Again, this third SMS bears traces of transidiomaticity (found longitudinally across time and informants), which includes non-elite multilingual practices (in Spanish and in English, too), phonetic spelling (as in *“felis”* for *“feliz”*, which means “happy”; or as in *“novides”* for *“navidades”*, for “Christmas”), and signs of xenoglossy involving aspects of morphology, again in greetings like *“buens noche”* (instead of

"*buenas noches*"). Shabbir's discourse is carefully planned and organised. It first includes two greetings and a very polite (yet unconventionally written) Spanish term of reference for the addressee ("*srts*", for "*srita*." or "Miss"), followed by the Christmas greeting, and, finally, by an accurate farewell establishing the term of address of the sender. Moreover, it makes frequent use of punctuation markers, using stops to divide up discursive units, and it also employs uppercase forms both for the respectful yet familiar term of address of the SMS receiver ("*Hermana*", "sister") and for his own name.

In fact, it could be argued that this SMS is hyper-polite — definitely the most polite in my small corpus. Shabbir refused to use informal contractions; for example, he knowingly typed the full form of "*hola*" for "hello", even though he explained that he knew about the frequent use of "*ola*" in the neighbourhood. Also, he included a polite face-saving term of reference which marks distance ("*srts*"), and, in short, seemed to be displaying many of his linguistic capitals in Spanish and in English, perhaps to be accountable to, and to present himself as a literate person in front of, an 'English teacher', the researcher. Thus, text message (7) resembles a very carefully typed, short formal letter rather than an everyday, rushed, informal SMS. This may demonstrate that transnational migrants' SMSes written in non-standard hybrid norms can also evoke different degrees of formality, as well as a wide range of politeness registers.

Besides, Shabbir chooses English, which indexes prestige in that part of South Asia (J.C. Paolillo 1996). I would argue that he does so as an *in-group sign of class distinction* and as a language choice which becomes an act of identity within one's transnational network in the neighbourhood. For Shabbir, whose wife in Kashmir is an English teacher and whose children are being educated in an English-medium elite private school, English indexes the class status of the wealthy intelligentsia in Pakistan and in Pakistani Kashmir (see T. Rahman 2002). This is re-appropriated, re-located and used in his transnational circuits of influence in the here and now, in front of other migrants who, like Naeem, may well know English but do not belong to any Pakistani circuits of power and have not experienced upward social mobility. In fact, this was supported by Shabbir's off-line practices. In the cyber-café, when I inquired about Pakistan, Shabbir would contradict Naeem's opinions by switching from Spanish into English, in turn positioning himself as a more legitimate knower of contemporary Pakistani and Kashmiri issues. The fact that he also emphasises his "Mughalness" in the SMS (which simultaneously indexes Muslimness, wealth and power; Farlex 2012) provides further evidence for this claim. Thus, the switch from Spanish to English may be "not just a contrast of languages but rather language styles selected for different kinds of identity work" (J. Androutsopoulos 2011: 294); in this case, that of self-ascribing a technoliterate, well-educated migrant identity.

3. The organisation of social life through fully-fledged hybrid “we-codes”

More often than not, the sort of entextualised multimodal hybrid Spanish which has been presented in the previous section is socially devalued, even by SMS receivers like Jenny who, when commenting on Moha’s SMS (1), interestingly told me (*i.e.* the “evaluative” researcher) that “*tendré que enseñarle a escribir*” (“I’ll have to teach him how to write”) in our interview. Jenny did so despite the fact that she, informally but consciously, engages in code-switching practices when she addresses her multilingual boyfriend. This is so because non-standard representations of language are always evaluated and framed within what is considered to be the standard, legitimate norm (A. Jaffe 2001, C. Thurlow 2006). Thus, to (normative) mainstream society, Moha’s, Naeem’s and Shabbir’s multilingual repertoires and ICT-mediated literacy practices index a *not-quite language* (S. Gal, 2006) employed by texters who are categorised as “deviant” or “incompetent” users of established literacy capitals, in this case in Spanish. In fact, these sorts of non-standard codes have derogatorily been called *immigrantese*, a label which, following C.A. Ferguson’s (1975) analysis of foreigner talk, publicly infantilises and pejoratively simplifies the migrants’ remarkable range of oral and written capitals in the host society’s languages (examples of the pejorative use of this term can be found in blog posts like Atari 2006, Axford 2009, or Topix 2009).

Against this prevailing regime of thought, I believe that a closer look at the social meanings or indexicalities of such linguistic and literacy practices demonstrates that transnational migrants have successfully mobilised this multimodal, flexible code for at least two reasons. On the one hand, their type-written hybrid Spanish, contingent on the non-elite multilingualism of given transnational communities, has become a powerful tool for migrants with differing linguistic capitals to access and self-distribute resources among group members, in their neighbourhood. That is, it has become their lingua franca or inter-group code for successful cross-cultural communication, one which becomes relevant and which gains social significance among users in the *immediate* social context in which it is circulated: in migrant-regulated networks and spaces like their own cybercafés.

As stated in the introduction, an ethnographically-grounded, comprehensive analysis of transnational SMS talk shall be contextualised and backed up by the non-elite multilingual practices and the vernacular literacies that the same informants under analysis employ in their immediate surroundings, as in the cybercafé of the neighbourhood. This is why I further support my claims by describing some of their hand-written practices, for I believe that many features of SMS talk are influenced by them. In the Vallès Occidental county, I found that the SMS transidiomatic practices that I have analysed so far, as well as the social meanings that these gain in transnational migration contexts, are also mobilised off-line in

hand-written room-for-rent advertisements, such as the one presented in Figure 1 below (photographed in the cybercafé of the same neighbourhood, on 6 August 2008).



Figure 1. Other non-elite hand-written practices with similar transidiomatic features, circulated by the transnational migrants of the neighbourhood. Picture taken by author

Figure 1 is a hand-written note mapped upon non-standard Spanish which shows traces of transidiomaticity; in particular, xenoglossy, as seen in the non-inflected bare infinitive “*alquilar*” (“rent”), on the one hand, and phonetic writing, as seen in the word “*abitacion*” for “*habitación*” (“room”), on the other. The street name also bears witness to the use of amalgamated spelling, as when “*señora de*” is written together (“*senorade*”). Since it was written by a middle-aged person of Moroccan origin (Moha’s friend), it could even be claimed that some of the Roman letters are a *transliteration* of the Arabic script (as exemplified by the three graphemes for “*n*”, or the grapheme for “*q*”, unconventionally written) — in fact I met several migrants who affirmed that they actually calqued the Spanish chunk from a model or sample, and just changed the street name and the contact number accordingly.

Despite the fact that the street name is somewhat confusing, what is relevant here is that the advertisement indicates that there is a room available, and that this is located in a specific area, such as for example near the public transport line. By looking at its graphic aesthetics, this language also indicates that such room is rented by a migrant person who belongs to the transnational networks of the neighbourhood; perhaps someone with similar mobile life trajectories and experiences. Nicely printed advertisements in standard Spanish inside the same cybercafé, by contrast, do not transmit the flavour of a non-elite multilingual voice, and tend to be attributed to what F. Barth (1969) calls an outsider or unknown *Other*. Besides, they also tend to index mistrust and, at times, even some sort of

“poshness”. This is how, via this code, the self-distribution of resources like rooms for rent is carried out from migrant hands to migrant hands.

Thus, on the other hand, transnational SMS practices, like off-line practices such as the ones presented in Figure 1, are also an index of the migrants’ acts of identity and a tool to gain a certain degree of *social agency* (A. Giddens 1984) within one’s own network; that which indexes non-elite multilingual living, group solidarity and affinity. In short, the transnational language of the migrants’ SMS is also a unique and particular voice or style, a “linguistic way of being” and a “linguistic way of (inter)acting” (C. Thurlow and K. Mroczek 2011b:xxx), whereby migrants successfully renegotiate their belonging and their entitlement to crucial city resources in current Catalan urban landscapes, both within mainstream society and inside their own networks.

Conclusion

In this paper I have first shown that, in order to fully understand the ever-changing virtual or ICT-mediated (and off-line) linguistic and literacy practices of transnational migrants, we need to go *beyond* the mere formal description of linguistic phenomena and analyse them from a more global, ethnographically-engaged and user-focused comprehensive, sociolinguistic approach. I have stressed that transnational migrants’ SMS talk is a unique window into key issues of concern within the sociolinguistics of globalisation; namely, into the still largely unexplored heteroglossic, multilingual repertoires and vernacular literacies of migrants, their mobility experiences inscribed in the language, and their re-negotiation of belonging to this increasingly interconnected world.

Secondly, I have argued that when focusing on flexible, hybrid linguistic practices we should pay close attention to the social meanings and the uses of the codes which are actually relevant and meaningful *for* the informants (C. Álvarez-Cáccamo 1998, P. Auer 1998; J. Blommaert 2008), avoiding *a priori* attributed or misguided preconceptions about the transnational citizens’ range of linguistic capitals and resources, and including their own readings and interpretations. This is crucial in order to problematize the widely publicised “immigrantese” discourses which pejoratively classify the migrants’ written practices that I have analysed here as “literacy gaps”, as “faulty momentary interlanguage stages”, or as mere “anecdotal pockets” resulting from the migrants’ “lack” of linguistic capitals or communicative resources.

In the third place, I have claimed that the migrants’ unique SMS practices have become the *norm* in transnational ICT-mediated communication. Migrants’ text messaging features are today rooted in transidiomaticity, including heterography

and xenoglossy, but neither the graphic/spelling difficulties nor the intrinsically transcodic marks in which they are embedded seem to hinder comprehension or prompt miscommunication. In fact, the opposite is true: linguistic informalisation, increased flexibility and hybridity have taken roots, and this endows migrants with a wide range of styles and with a command of several degrees of formality and informality employed for performativity and discourse-stylisation purposes, giving way to the ICT-mediated transnational identities of the global era. Hence, transnational SMSes may bear old and new identity features, including, for example, working-class alignment via the use of the Spanish youngsters' anti-normative textese language (as in the case of the informant Moha), or the presentation of the self as a member of a given elite via the elaboration of extremely polite terms of address (as in Shabbir's SMS).

This finding challenges the assumption that a certain level of competence is a prerequisite for successful ICT-mediated multilingual communication, and demonstrates that transnational SMS users can actually employ whatever linguistic features they have at their disposal and achieve all their communicative aims through their own means, regardless of how well they command and know the languages involved (J.N. Jørgensen *et al.* 2012; J. Androutsopoulos 2012). In other words, transnational SMSes show that migrants have the capacity to insert themselves into contemporary communicative and literacy regimes and to overcome or even *subvert* their linguistic marginalisation successfully. Thus, they self-cater, compete for, and gate-keep the access to crucial material and symbolic resources via the appropriation of particular ICT-mediated codes which today have become the migrants' *inter-group counter-hegemonic linguae francae*.

A final word of caution is needed concerning the migrants' successful organisation of their transnational lives and prospects through these codes, though. While they may be empowering and liberating in the ways that I have just highlighted, the emergent non-elite allochthonous ICT-mediated talk that I observed unfolds within the macro linguistic hierarchies and the sociolinguistic regimes in which current transnational citizens are forcibly socialised today. In the Catalan context, this means that the migrants under study, who massively typed primarily in Spanish, seem to reproduce the prevailing mainstream Spanish monolingual norm as well as a monoglossic frame of mind, to the detriment of their own linguistic capitals, paradoxically contributing to the silencing of already minoritised languages and, thus, to the reproduction of social inequality through ICT-mediated talk.

References

- Alonso, Andoni; Pedro J. Oiarzabal (eds.). 2010. *Diasporas in the New Media Age. Identity, Politics, and Community*. Reno, Nevada: University of Nevada Press.
- Álvarez-Cáccamo, Celso. 1998. From 'Switching Code' to 'Code Switching': Towards a Reconceptualisation of Communicative Codes. In *Code-Switching in Conversation: Language, Interaction and Identity*, P. Auer (ed.), 29–51, London/New York: Routledge.
- Androutsopoulos, Jannis. 2006a. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10(4): 419–438, Oxford: Blackwell.
- Androutsopoulos, Jannis. 2006b. Multilingualism, diaspora, and the Internet: Codes and identities on German-based diaspora websites. *Journal of Sociolinguistics* 10(4): 520–547, Oxford: Blackwell.
- Androutsopoulos, Jannis. 2007. Language Choice and Code Switching in German-based Diasporic Web Forums. In *The Multilingual Internet. Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 340–361, Oxford: Oxford University Press.
- Androutsopoulos, Jannis. 2011. From Variation to Heteroglossia in the Study of Computer-Mediated Discourse. In *Digital Discourse. Language in The New Media*, C. Thurlow, K. Mroczek (eds.), 277–298, Oxford: Oxford University Press.
- Androutsopoulos, Jannis. 2012. Network(ed) choices: Towards understanding multilingualism in social media. BAAL/CUP Seminar: Language and Social Media, The University of Leicester, 26–27 April 2012.
- Anis, Jacques. 2007. Neography: Unconventional Spelling in French SMS Text Messages. In *The Multilingual Internet. Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 87–115, Oxford: Oxford University Press.
- Appadurai, Arjun. 1996. *Modernity at Large. Cultural Dimensions of Globalization*. Minneapolis: University of Minnesota Press.
- Area Moreira, Manuel; Begoña Gros Salvat; Miguel A. Marzal García-Quismondo. 2008. *Alfabetizaciones y tecnologías de la información y la comunicación*. Madrid: Síntesis.
- Atari. 2006. *Atari Community Forums*, blog. Accessed from <<http://ataricomunity.com/forums/>>, [10 October 2009].
- Auer, Peter (ed.). 1998. *Code-Switching in Conversation: Language, Interaction and Identity*. London/New York: Routledge.
- Axford, Emily. 2009. 'WWJD?'; e.e. cumming all over your titties, blog post, 22 January. Accessed from <<http://emilyaxford.wordpress.com/>>, [10 October 2009].
- Baron, Naomi S. 2008. *Always on: Language in an Online and Mobile World*. Oxford/New York: Oxford University Press.
- Barth, Frederick (ed.). 1969. *Ethnic Groups and Boundaries: The Social Organisation of Cultural Difference*. Boston: Little, Brown and Company.
- Blommaert, Jan. 2003. Commentary: Asociolinguistics of globalization. *Journal of Sociolinguistics* 7(4): 607–623.
- Blommaert, Jan. 2008. *Grassroots Literacy. Writing, Identity and Voice in Central Africa*. London/New York: Routledge.
- Blommaert, Jan. 2010. *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press.
- Blommaert, Jan. 2012. *Chronicles of Complexity: Ethnography, Superdiversity, and Linguistic Landscapes*. Tilburg Papers in Cultural Studies 29. <<http://www.tilburguniversity.edu/research/institutes-and-research-groups/babylon/tpcs/paper29.pdf>>.

- Blommaert, Jan; Ben Rampton. 2011. Language and superdiversity: A position paper. *Working Papers on Urban Language and Literacies* 70. <<http://www.kcl.ac.uk/innovation/groups/ldc/publications/workingpapers/70.pdf>>.
- Borochofsky-Bar-Aba, Esther; Yafit Kedmi. 2010. The nature of SMS discourse: The case of Hebrew. *Folia Linguistica* 44(1): 1–30.
- Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*. Cambridge/New York: Cambridge University Press.
- Bourdieu, Pierre. 1990. *The Logic of Practice*. Stanford: Stanford University Press.
- Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Cambridge, Massachusetts: Harvard University Press.
- Cameron, Deborah. 2000. *Good to Talk? Living and Working in a Communication Culture*. London/California/New Delhi: Sage.
- Castells, Manuel. 2000 [1996]. *The Rise of the Network Society*. Oxford: Blackwell.
- Castells, Manuel; Mireia Fernández-Ardévol; Jack Linchuan Qiu; Araba Sey. 2007a. *Mobile Communication and Society. A Global Perspective*. Cambridge: MIT Press.
- Castells, Manuel; Imma Tubella; Teresa Sancho; Meritxell Roca. 2007b. *La transició a la societat xarxa*. Barcelona: Ariel.
- Chipchase, Jan. 2008. Reducing Illiteracy as a Barrier to Mobile Communication. In *Handbook of Mobile Communication Studies*, J.E. Katz (ed.), 79–89, Cambridge: MIT Press.
- Coupland, Nik. 2003. Introduction: Sociolinguistics and globalisation. *Journal of Sociolinguistics* 7(4): 465–472.
- Coupland, Nik (ed.). 2010. *The Handbook of Language and Globalization*. Oxford: Wiley-Blackwell.
- Crystal, David. 2008. *Txtng: The Gr8 Db8*. Oxford/New York: Oxford University Press.
- Danet, Brenda; Susan C. Herring. 2007. Introduction: Welcome to the Multilingual Internet. In *The Multilingual Internet. Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 1–39, Oxford: Oxford University Press.
- Deumert, Ana; Sibabalwe Oscar Masinyana. 2008. Mobile language choices — The use of English and isiXhosa in text messages (SMS) Evidence from a bilingual South African sample. *English World-Wide* 29(2): 117–147.
- Duranti, Alessandro. 1997. *Linguistic Anthropology*. New York: Cambridge University Press.
- Fairclough, Norman. 2006. *Language and Globalization*. London/New York: Routledge.
- Farlex. 2012. Mogul. *The Free Dictionary by Farlex*. <<http://www.thefreedictionary.com/Mogul>>.
- Ferguson, Charles A. 1975. Toward a Characterization of English Foreigner Talk. *Anthropological Linguistics* 17(1): 1–14.
- FOBSIC. 2010. *Enquesta sobre l'equipament i l'ús de les Tecnologies de la Informació i la Comunicació (TIC) a les llars de Catalunya (2010). Volume II. Usos TIC individuals*. Fundació Observatori per la Societat de la informació a Catalunya and Institut d'Estadística de Catalunya. <<http://www.fobsic.cat/>>.
- Gal, Sue. 2006. Migration, Minorities and Multilingualism: Language Ideologies in Europe. In *Language Ideologies, Policies and Practices*, C. Mar-Molinero, P. Stevenson (eds.), 13–27, Basingstoke: Palgrave.
- Generalitat de Catalunya. 2006. *Estatut d'Autonomia de Catalunya*. Barcelona: BIGSA Indústria Gràfica.
- Giddens, Anthony. 1984. *The Constitution of Society. Outline of the Theory of Structuration*. Cambridge: Polity Press.

- Goggin, Gerard. 2006. *Cell Phone Culture. Mobile Technology in Everyday Life*. London/New York: Routledge.
- Grinter, Rebecca E.; Margery Eldridge. 2003. Wan2tlk?: Everyday text messaging. *CHI '03 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. <<http://www.cc.gatech.edu/~keith/classes/ubicomplexity/pdfs/crit/grinter-sms.pdf>>.
- Gumperz, John J. 1982. *Discourse Strategies*. Cambridge: Cambridge University Press.
- Gumperz, John J. 1986. Interactional Sociolinguistics in the Study of Schooling. In *The Social Construction of Literacy*, Jenny Cook-Gumperz (ed.), 45–68, Cambridge: Cambridge University Press.
- Heller, Monica. 2007. (ed.). 2007. *Bilingualism: A Social Approach*. London: Palgrave Macmillan.
- Heller, Monica. 2010. *Paths to Post-Nationalism. A Critical Ethnography of Language and Identity*. New York: Oxford University Press.
- Horst, Heather A.; Daniel Miller. 2006. *The Cell Phone: An Anthropology of Communication*. Oxford: Berg.
- Howard, Philip N. 2002. Network ethnography and the hypermedia organization: New media, new organizations, new methods. *New Media & Society* 4(4): 550–574.
- Idescat. 2011. Població segons nacionalitat. Institut d'Estadística de Catalunya. <<http://www.idescat.cat/>>.
- Jacquemet, Marco. 2005. Transidiomatic practices: Language and power in the age of globalization. *Language & Communication* 25(3): 257–277.
- Jacquemet, Marco. 2010. Language and transnational spaces. In *Language and Space: An International Handbook of Linguistic Variation*, P. Auer, J.E. Schmidt (eds.), 50–69, New York: Mouton de Gruyter.
- Jaffe, Alexandra. 2000. Introduction: Non-standard orthography and non-standard speech. *Journal of Sociolinguistics* 4(4): 497–513.
- Jørgensen, J. Normann; Rosita Rindler-Schjerve; Eva Vetter. 2012. Polylingualism, Multilingualism, Plurilingualism. *A Toolkit for Transnational Communication in Europe*. <<http://www.toolkit-online.eu/docs/polylingualism.html>> [accessed 18 April 2012].
- Kahn, Richard; Douglas Kellner. 2005. Reconstructing technoliteracy: A multiple literacies approach. *E-learning* 2(3): 399–423.
- Katz, James E. (ed.). 2008. *Handbook of Mobile Communication Studies*. Cambridge: MIT Press.
- Katz, James E.; Mark Aakhus. 2002. *Perpetual Contact. Mobile Communications, Private Talk, Public Performance*. Cambridge: Cambridge University Press.
- Kress, Gunter; Theo van Leeuwen. 2006 [1996]. *Reading Images. The Grammar of Visual Design*. London/New York: Routledge.
- Kroskrity, Paul V. (ed.). 2000. *Regimes of Language. Ideologies, Politics, and Identities*. Santa Fe/ New Mexico/Oxford: School of American Research Press and James Currey Ltd.
- Lam, Eva W.S. 2009. Multiliteracies on instant messaging in negotiating local, translocal, and transnational affiliations: A case of an adolescent immigrant. *Reading Research Quarterly* 44(4): 377–397.
- Lam, Eva W.S.; Enid Rosario-Ramos. 2009. Multilingual literacies in transnational digitally mediated contexts: An exploratory study of immigrant teens in the United States. *Language and Education* 23(2): 171–190.
- Lam, Eva W.S.; Doris Warriner. 2012. Transnationalism and literacy: Investigating the mobility of peoples, languages, texts, and practices in the context of migration. *Reading Research Quarterly* 47(2): 191–215.

- Ling, Rich. 2004. *The Mobile Connection: The Cell Phone's Impact on Society*. San Francisco, California: Morgan Kaufmann.
- Ling, Rich. 2005. The Sociolinguistics of SMS: An Analysis of SMS Use by a Random Sample of Norwegians. In *Mobile Communications. Re-negotiation of the Social Sphere*, R. Ling, P. Pedersen (eds.), 335–349, London: Springer.
- Martín-Rojo, Luisa; Clara Molina. 2012. Madrid multilingüe: Lenguas pa' la citi. Excelencia UAM. Artículos y Reportajes. <http://www.uam.es/ss/Satellite/es/1242652961025/1242664605633/articulo/articulo/Madrid_multilingue:_Lenguas_pa%E2%80%99_la_citi.htm>.
- Palfreyman, David; Muhamed Al Khalil. 2007. "A Funky Language for Teenzz to Use": Representing Gulf Arabic in Instant Messaging. In *The Multilingual Internet. Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 43–63, Oxford: Oxford University Press.
- Panagakos, Anastasia N.; Heather A. Horst. 2006. Return to Cyberia: Technology and the social worlds of transnational migrants. *Global Networks* 6(2): 109–124.
- Paolillo, John C. 1996. Language choice on soc.culture.punjab. *Electronic Journal of Communication* 6(3). <<http://www.cios.org/EJCPUBLIC/006/3/006312.HTML>>.
- Pennycook, Alastair. 2007. *Global Englishes and Transcultural Flows*. London/New York: Routledge.
- Peuronen, Saija. 2011. "Ride Hard, Live Forever": Translocal Identities in an Online Community of Extreme Sports Christians. In *Digital Discourse. Language in The New Media*, C. Thurlow, K. Mroczek (eds.), 154–176, Oxford: Oxford University Press.
- Rahman, Tariq. 2002. *Language, Ideology and Power: Language-Learning among Muslims of Pakistan and North India*. Karachi/Oxford: Oxford University Press.
- Rheingold, Howard. 2002. *Smart Mobs. The Next Social Revolution. Transforming Cultures and Communities in the Age of Instant Access*. Cambridge, Massachusetts: Perseus.
- Robledo, Juanjo. 2008. 'Generación-i. La transición multicultural', *El País semanal*, weekly magazine, 6 July. Accessed from <<http://www.elpais.com/>> [5 June 2009].
- Ros, Adela; Elisabet González; Antoni Marín; Papa Sow. 2007. Migration and information flows: A new lens for the study of contemporary international migration. *Working Paper Series WP07-002*. <<http://www.uoc.edu/>>.
- Sabaté i Dalmau, Maria 2012a. 'The official language of Telefónica is English': Problematising the Construction of English as a lingua franca in the Spanish telecommunications sector. *Atlantis* 34(1): 133–151.
- Sabaté i Dalmau, Maria. 2012b. Aportaciones de la etnografía de red al estudio de un locutorio: hacia un cambio de paradigma metodológico. In *La sociolingüística con perspectiva etnográfica en el mundo hispano. Nuevos contextos, nuevas aproximaciones*, E. Codó, A. Patiño, V. Unamuno. Special Issue in *Spanish in Context* 9 (2): 191–218.
- Sabaté i Dalmau, Maria. 2012c. Fighting exclusion from the margins: Locutorios as sites of social agency and resistance for migrants. In *Language, Migration, and Social Inequalities*, A. Duchêne, M. G. Moyer, C. Roberts (eds.), 248–271, Bristol: Multilingual Matters.
- Sheller, Mimi; John Urry. 2006. The new mobilities paradigm. *Environment and Planning A* 38: 207–226.
- Slater, Don; Daniel Miller. 2000. *The Internet: An Ethnographic Approach*. Oxford: Berg.
- Solé, Carlota; Sònia Parella; Leonardo Cavalcanti. 2007. *L'empresariat immigrant a Espanya*. Col·lecció Estudis Socials 21. Barcelona: Fundació La Caixa.
- Solé, Carlota; Sònia Parella; Leonardo Cavalcanti (eds.). 2008. *Nuevos retos del transnacionalismo en el estudio de las migraciones*. Ministerio de Trabajo e Inmigración. Subdirección General de Información. Madrid: Grafo.

- Tagg, Caroline. 2012. *The Discourse of Text Messaging. Analysis of SMS Communication*. London: Continuum.
- Thurlow, Crispin. 2006. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication* 11(3): 667–701.
- Thurlow, Crispin; Alex Brown. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*. <<http://extra.shu.ac.uk/daol/>>.
- Thurlow, Crispin; Kristine Mroczek (eds.). 2011a. *Digital Discourse. Language in The New Media*. Oxford: Oxford University Press.
- Thurlow, Crispin; Kristine Mroczek. 2011b. Fresh Perspectives on New Media Sociolinguistics. In *Digital Discourse. Language in The New Media*, C. Thurlow, K. Mroczek (eds.), xix–xliv, Oxford: Oxford University Press.
- Topix. 2009. 'Defend to end!'. *Topix*, blog post, 13 February. Accessed from <<http://www.topics.com/forums/>> [10 October 2010].
- Torres i Vilatarsana, Marta; Lluís Payrató. 2003. El català dels joves en els xats, correus electrònics i missatges a mòbils: una nova varietat col·loquial? In *Llengua, Societat i Ensenyament (Volum III)*, Vicent Martines (ed.), 5–27, Institut Universitari de Filologia Valenciana Symposia Philologica. Alacant: Tàbula.
- Vertovec, Steven. 2001. Transnationalism and identity. *Journal of Ethnic and Migration Studies* 27(4): 573–582.
- Vertovec, Steven. 2006. The emergence of super-diversity in Great Britain. *Centre on Migration, Policy and Society Working Papers* 25. Oxford: Oxford University Press.
- Vertovec, Steven. 2009. *Transnationalism*. London/New York: Routledge.
- Woolard, Kathryn A. 1989. *Double Talk: Bilingualism and the Politics of Ethnicity in Catalonia*. Stanford, California: Stanford University Press.
- Woolard, Kathryn A. 2003. 'We Don't Speak Catalan Because We Are Marginalized'. Ethnic and Class Connotations of Language in Barcelona. In R. Blot (ed.), 85–103, *Language and Social Identity*. Westport, Connecticut: Praeger.

Summary

From the field of the sociolinguistics of globalisation, this article investigates the linguistic features of transnational SMS talk, focusing on the heteroglossic and hybrid multilingual text messaging practices and the ICT-mediated vernacular literacies of a very heterogeneous small group of migrants establishing transnational networks in the outskirts of Barcelona. It shows that migrants employ highly flexible, non-elite *linguae francae* or “we-codes” for successful inter-group communication which are based on heterography, orality, anti-standardness and transidiomaticity. It also explores the social indexicalities of such SMS practices, and claims that, against a highly ideologised discursive regime which classifies them as “faulty” or “deviant”, transnational migrants' text messages offer an insight into how these highly mobile citizens attain the necessary degree of social agency to unfold their many transnational identities, re-negotiate their belonging and entitlement to host-society resources, and manage to organise their life trajectories and prospects largely successfully.

Negation marking in French text messages

Elisabeth Stark
Universität Zuerich

Introduction

In contemporary Standard French, the sentential negation is expressed, with finite verb forms, by a preverbal clitic, *ne*, and postverbal negative adverbs (1 a-b) or negative pronouns (1 c):

- (1) a. *Je ne sais pas parler allemand.*
(‘I do not speak German.’)
- b. *Je ne sais plus parler allemand.*
(‘I do not speak German any longer.’)
- c. *Je ne connais personne.*
(‘I do not know anybody.’)

French is a *negative concord* language (cf. M. Haspelmath 1997, P. Larrivée 2004) in that the co-occurrence of the negative clitic *ne* and non-clitic postverbal negative elements like *pas* (‘not’), *plus* (‘any more’) or *personne* (‘nobody’) does not result in double negation.

In non-standard French however, *ne* is frequently dropped. The literature does not agree on specific circumstances and triggers of this drop, but it can be stated that *ne*-drop is indicative of non-standard uses of French, typically in everyday conversational contexts.

As the vast majority of text messages are considered prototypical instances of casual computer-mediated communication (CMC, cf. J.B. Walther 1996), we would expect a reasonably high rate of *ne* deletion in this rather ‘standard-free’ form of communication (cf. C. Dürscheid 2002, C. Dürscheid and E. Stark 2011, E. Stark 2011). Yet, text messages are graphic, *i.e.* written, they are not phonic in nature, and the normative influence of writing rules learned at school may inhibit the spontaneous spelling of the corresponding natural, phonic utterances (besides other factors like asynchronicity, see C. Dürscheid 2002 and below). Text messages can be functionally characterized as a “technology of sociability”, showing “immediacy and intimacy” and “metapragmatic awareness” by their authors

(cf. Thurlow/Poff 2013). In their seminal overview of the “nature of texting”, C. Thurlow and M. Poff (2013) mention three “maxims of texting”, *i.e.* “brevity and speed; paralinguistic restitution; phonological approximation” (C. Thurlow and M. Poff 2013: 176). These all have to be taken into account when dealing with this kind of linguistic data and are specific to this form of communication as opposed to electronic mail, for example.

To our knowledge, there have been no empirical studies that investigate negation marking in French text messages until now. This may be due to a lack of authentic large-scale corpora of French text messages until the international corpus project *sms4science* (cf. www.sms4science.org) started in Belgium in 2004 and spread to four other French-speaking regions in the world (Réunion Island, Switzerland, Québec and France, cf. Section 2).

Fortunately, there are some important studies by Rémi van Compernelle (e.g. 2008 and 2010) on *ne* realization in French chats, *i.e.* online instant messaging. In the new communication forms of computer-mediated communication (= CMC in general, text messages in particular), there are two things that deserve special attention: firstly, the relatively new combination of rather informal, non-official, but *graphic* discourse, which has given rise to important modifications of traditional variationist approaches to ‘oral’ or ‘spoken’ vs ‘written’ language use; secondly, the new factor of synchronicity, *i.e.* the distinction between communication forms where two people can directly interact, interrupt each other and follow each other’s language production ‘online’ (synchronous) and communication forms where this is not possible (asynchronous).

The degree to which CMC can be associated with speech and/ or writing, however, often depends on its level of synchronicity [...]. On the one hand, synchronous — and quasi-synchronous [...] — forms of CMC (e.g., chat and instant messaging) are generally regarded as more conversational forms of communication, thus leading researchers to think of them as *text-based forms of ‘spoken’ language*. On the other hand, asynchronous CMC (e.g., e-mail, discussion fora, listservs, and web pages) have *come to be associated with the written language*, because their asynchronous nature allows for virtually unlimited editing and revision before a message is sent [...]. Their findings suggest a close alignment between synchronous chat and informal speech. *On the other hand, asynchronous forms of CMC tend to align more closely with the written language.* (R. van Compernelle 2010: 448f., our emphasis)

We think that it is necessary to distinguish systematically the *conception* of a message, *i.e.* its lexical and morphosyntactic make-up (and, in the phonic modality, its phonetic and phonological make-up) according to the degree of formality of the situation (e.g. an exam or a police interrogation vs a casual conversation among friends, cf. P. Koch and W. Oesterreicher 2001, 2011) from the purely *material*

aspect of its encoding, called *medium* in what follows (graphic vs phonic, in the sense of “spoken language” in the quote above). Only then are we able to see that what has been considered typically ‘oral’ or ‘spoken’, like *ne* deletion in French, may be due to at least two different factors — *conception* or *medium* (cf. P. Koch and W. Oesterreicher 2011). The difference between and the independency of conception and medium, which are sometimes both called “written” and “spoken”, can be seen from the following figure:

- (2) Conception and medium of a message following P. Koch and W. Oesterreicher 2011

		CONCEPTION	
		spoken	written
MEDIUM	graphic	<i>faut pas le dire</i>	<i>il ne faut pas le dire</i>
	phonic	[fopɑl'dir]	[ilnəfopɑlɑ'dir]

We can see from this figure that in a communication with a “spoken conception”, e.g. among friends at home having dinner together, the expletive subject *il* and the negation particle *ne* can be dropped in an utterance such as *faut pas le dire* (‘you should not say that/it’). The same is true in a text message exchanged between friends. These elements may similarly be omitted in other written forms, depending on the addressee, the relationship with him/her and other communicative parameters. On the other hand, in a more formal situation such as an oral exam, one would most probably use *il* and *ne*, just as in a written exam. That means that the absence of *il* and *ne* is an indicator of the informal *conception* of a message and not so much of its modality or *medium* (i.e. sounds = *phonic* vs letters = *graphic*). Figure (2) shows the principled independency of these two dimensions, but there are prototypical affinities (such as informal and phonic and formal and graphic), which are altered by the arrival of CMC.

With the new communication forms of CMC at hand, we are able — almost for the first time — to analyse large amounts of graphic material exchanged frequently in unofficial, casual situations. This analysis may provide new insight into variation phenomena of all sorts. In what follows, we use this new data in an analysis of *ne* retention and *ne* deletion in a corpus of French text messages; something that, to our knowledge, has not been done before. Text messages are considered a form of asynchronous CMC and thus closer to “written language” (cf. R. van Compernelle 2010), but they show a considerable number of features of non-standard French.

The paper is structured as follows: Section 1 gives an overview of the main findings concerning *ne* deletion, both in historical and contemporary French, with a focus on the most important language-external and language-internal triggers

for *ne* deletion. We will formulate four general hypotheses which will guide our own corpus analysis and help interpret our findings. Section 2 presents our data, the Swiss corpus of text messages, *sms4science.ch*, and our methodology. Section 3 reviews the results of the corpus analysis. It focuses on three language-internal factors for *ne* deletion, *i.e.* subject type (3.1), second negative element (3.2), and different verb forms (3.3). Section 4 is a discussion of our findings in the context of wider research. It will show that our findings confirm overall internal regularities of *ne* deletion, but to a different degree compared to phonic material. Thus, our findings support R. van Compernelle's (2008, 2010) hypothesis about the special "text-based" or more precisely graphic character of CMC. Section 5 summarizes the main results in a brief conclusion.

1. State of the art and hypotheses

Whereas traditional reference grammars (cf. M. Grevisse and A. Goosse, 1993: 1457–1458, M. Riegel, J.-C. Pellat and R. Rioul 1994: 411) and textbooks of French describe and recommend the bipartite negation *ne...pas/plus/personne* etc. with finite verbs (see examples under (1)), it is well known that the deletion of the first part of sentential negation in contemporary French is a very widespread phenomenon (for illustration, see the examples in the left part of the table under (2)):

Ne, souvent omis à l'oral, perd de son importance en français moderne [...].
(M. Riegel, J.-C. Pellat and R. Rioul 1994: 415)

It is even sometimes considered a genuine sociolinguistic variable (cf. W. Ashby 1976, 1981 and A. Coveney 2002) or at least a variety indicator of rather low registers in French (cf. A. Dufter and E. Stark 2002; A. Dufter and E. Stark 2007 for this view on French *ne* omission, also in a diachronic approach).

As for the diachronic development, it is assumed generally that *ne* has seen continuously less use since at least the 19th century (cf. e.g. W. Ashby 2001; N. Armstrong and A. Smith 2002; A.-B. Hansen and I. Malderez 2004; F. Martineau and R. Mougeon 2003). A. Dufter (2012) shows a slight increase in *ne* deletion rates in graphic literary and philosophical texts, where it remains marginal and most probably represents a variety indicator for low registers deliberately chosen by the respective authors:

(3) *ne* deletion in *Frantext*¹ between 1700 and 1999 (cf. A. Dufter 2012)

	1700–1749	1750–1799	1800–1849	1850–1899	1900–1949	1950–1999
– <i>ne</i>	1	10	76	246	2422	5668
Tot. neg	7396	14598	18631	20479	36579	34902
%– <i>ne</i>	<0.1%	<0.1%	0.4%	1.2%	6.6%	16.2%

Based on Canadian private writings like letters and diaries, F. Martineau and R. Mougeon (2003) propose that the 19th century marks the beginning of a radical *ne* drop in Canadian French, and F. Martineau (2009) even observes a 20% rate of *ne* deletion in a corpus of Canadian private letters from the 18th century. The quasi-phonetic transcription of authentic utterances of Louis XIII and people surrounding him that were gathered in the *Journal d'Héroard* from the 17th century (cf. G. Ernst 1985) show quite high rates of *ne* deletion (cf. A. Dufter and E. Stark 2007); thus, *ne* is dropped in the *Journal d'Héroard* (comprising the utterances of the young *Dauphin* as well as those of adults) with lexical subjects in 62.3% of all cases, and with pronominal subjects in 62.8%.

Corpus studies of phonic French material from the 20th and 21st century since 1940 show, with only three exceptions (W. Ashby 1976, A.-M. Diller 1983, J. Pohl 1968), *ne* deletion rates of above 50%, more precisely between 60% and 98% (cf. N. Armstrong 2002, W. Ashby 1981, W. Ashby 2001, A. Coveney ²2002, B.-B. Fonseca-Greber 2007, A.-B. Hansen and I. Malderez 2004, to cite only the most important ones). The particularly high rates of *ne* deletion typically show up in spontaneous speech. Yet, in speech used in radio broadcasts, *ne* retention is remarkably high even in the late 1990s (cf. N. Armstrong and A. Smith 2002 with only about 30% of *ne*-drop). This is a clear indicator that the *ne* drop depends on the degree of 'informality'; with a public audience, *ne* is retained more often.

For our work, the studies of R. van Compernelle (2008, 2010) on French on-line chats are very important. This research is based on another form of computer-mediated communication, with the following overall results:

In the case of *ne* use, the data analysed in this study suggest that the discourse of IRC [= Internet Relay Chat] is similar to that of everyday conversational French in that non-standard forms and structures, such as single-particle negation (*i.e.*, *ne* deletion), are not only common, but are part of the complex of sociolinguistic norms established by online communities. However, the data also indicate that the variable use of *ne* occurs somewhat differently in chat discourse as compared to previous studies of everyday conversational French, which may be due — at least

1. *Frantext*, cf. www.frantext.fr, is a data base of *graphic* and also mainly formal, *written* French, composed by more than 4000 authentic French texts from literature, philosophy, and science, dating from 1180 to 2009.

in part — to the text-based nature of synchronous online communication. (R. van Compernelle 2008: 318)

In fact, the literature identifies two groups of triggers for *ne* deletion: external and internal, *i.e.* linguistic. As for external factors, we will discuss two types in what follows: regional and stylistic-pragmatic factors (for example, the influence of the situation such as the communication partners' relationship). As is shown in (4), both are considered relevant (together with age and education/social class, which are not taken into account in what follows):

(4) External triggers for *ne* deletion

	+ <i>ne</i> -	- <i>ne</i>
Region (cf. A. Coveney ² 2002, B.-B. Fonseca-Greber 2007)	Southern France	Canada, Switzerland, Northern France
Situation type (cf. N. Armstrong 2002, N. Armstrong and A. Smith 2002, W. Ashby 1981, A. Coveney ² 2002, C. Meisner 2010)	official, radio	private, <i>face-to-face</i>

In fact, in Canadian French, *ne* is no longer considered a morphosyntactic element available to most native speakers (cf. e.g. J. Auger 1994, and R. van Compernelle 2008, 2010 for European French *vs* Canadian French online chats). Strong tendencies towards *ne* deletion can also be observed in Northern France, and in Switzerland, where our data stem from (see below, Section 2). Switzerland is also strongly affected by *ne* drop (2.5% *ne* retention in B.-B. Fonseca-Greber's (2007) data, phonic everyday conversations). In addition, for the first time, C. Meisner (2010) shows systematically that stylistic-pragmatic factors are relevant for *ne* drop. C. Meisner (2010) compares two subcorpora taken from the phonic reference corpus for Romance languages, *C-ORAL-ROM* (cf. E. Cresti and M. Moneglia 2005). The first, made up of recordings from political debates and university lectures, shows a *ne* retention rate of 64%, whereas the second, consisting of private conversations between family members and friends, shows only 4% *ne* retention.

As for internal, structural triggers, the literature mainly discusses the following (besides frequent collocations, simple *vs* compound verbal tenses, matrix *vs* subordinate clauses, intervening adverbs, phonological context, not taken into account in what follows):

(5) Internal triggers for *ne* deletion

	+ <i>ne</i>	– <i>ne</i>
Subject type (cf. N. Armstrong and A. Smith 2002: 31, A. Coveney ² 2002: 73, A.-B. Hansen and I. Malderez 2004: 21, A. Dufter and E. Stark 2007: 122–124, C. Meisner 2010: 1949, R. van Compernelle 2008)	non clitic (proper noun <i>Paul</i> , full noun phrase, stressed pronoun, <i>nous, vous</i> , relative or indefinite pronoun <i>qui, quelqu'un</i>)	clitic (<i>je, tu, il(s)</i>)
Subject doubling (cf. A.-B. Hansen and I. Malderez 2004: 21)	no doubling (XP ² <i>ne</i> V Neg)	doubling (XP CL V Neg)
Form and frequency of postverbal negative element (cf. N. Armstrong and A. Smith 2002: 37, Ashby 1981: 682, A.-B. Hansen and I. Malderez 2004: 22, M.-L. Moreau 1986: 137–160)	<i>que</i>	<i>pas</i>
Non-finite verb forms / imperative vs indicative verb forms (cf. W. Ashby 1976 : 123, W. Ashby 1981 : 679, C. Meisner 2010 : 1949)	infinitive / imperative	indicative

Ne retention is considerably higher with lexical subjects (proper nouns included; cf. example (6) a below) and with *nous* (at least 50% in the different studies, also in the most relevant one for us, R. van Compernelle 2008). Lexical subjects show *ne* retention rates ranging from 92.4% in the radio broadcasts of N. Armstrong and A. Smith (2002) and 56.4% in casual spoken French in A.-B. Hansen and I. Malderez (2004). Still smaller percentages of *ne* retention are found with the relative pronoun *qui*, the demonstrative *ça* and second person plural *vous*, followed again by third person clitics *il(s)* and *elle(s)*. The lowest rates of *ne* retention are found with *je, tu, ce*, and *ne* is categorically missing with doubled subjects (e.g. *ma mère elle veut pas, moi je peux pas* etc.). The following examples, taken from our corpus (see Section 2), illustrate these findings:

- (6) a. *J'espère que ta soirée ne s'est pas trop mal passée* (lexical subject)
(‘I hope that your evening was not that bad’)
- b. [...] *vous vous êtes pas perdues? à toute, profite bien* [...] (‘Didn’t you get lost? See you, enjoy it’) (*ne*-drop with *vous*_{2,Pl})
- c. *Elles viendront pas je pense* [...] (*ne*-drop with *elles*_{3,Pl,F})
(‘I don’t think they will come’)
- d. *jpeu pas non plus. Mardi ché pas ?!* (*ne*-drop with *je*_{1,Sg})
(‘I cannot come either. Tuesday, don’t know?’)

- e. *la plupart des gens ils connaissent pas [...]* (*ne*-drop with doubled subject)
(‘the majority of people don’t know’)

R. van Compernelle (2008) shows an overall *ne* realization rate of 16.1% in European French-language online chat, which parallels overall tendencies in contemporary spoken European French. He also identified subject type as the most important linguistic factor that triggered *ne* retention (59.6% with lexical subjects, only 12% with pronominal ones). He found relatively high *ne* retention rates with non-finite and imperative verb forms (19.2% of *ne* retention). For the latter language-internal factor, see the following examples from our corpus ((7) a shows *ne pas* in front of an infinitive and (7) b shows *ne...pas* negating two present participles, both in accordance with Standard French regularities):

- (7) a. *J'aurai bien voulu resté mais je préfère ne pa etre bien à la maison !!*
(‘I would have loved to stay but I prefer to feel bad at home’)
b. *N'arrivant pas à vous joindre et n'ayant pas de nouvelles [...]*
(‘As I cannot contact you and do not have any news from you’)

However, the relevance of the non-clitic negative element (*pas*, *plus*, *personne* etc.) could not be shown (see below). R. van Compernelle 2010 presents an analysis of Canadian chats and finds considerably lower figures (overall *ne* retention of only 5.1%), which parallels findings on spoken Canadian French (see above). It is important to highlight that in both cases, percentages of *ne* retention are slightly higher than in phonic corpora, perhaps due to the written, graphic modality of CMC, while the observed language-internal regularities mirror the general tendencies. Furthermore, R. van Compernelle (2008: 335) states that a fine-grained analysis of subject clitics and their phonological context shows sensitivity of *ne* for preceding and following phonemes. It is retained between vowel and consonant or consonant and vowel (see examples, (6) a above), whereas it is dropped between two consonants (see example (6) c above) and even more so between two vowels (e.g. <tu n'es pas gentil> — <t'es pas gentil>, ‘you are not nice’). This is reminiscent of typical writing strategies in CMC such as “letter-homophones” (cf. C. Thurlow 2003: <t>, [tɛ], ‘[letter] t’ for nearly homophonic <t'es>, [tɛ], ‘you are’), and mirrors affixation tendencies of the subject clitics in contemporary French (cf. J. Culbertson 2010, E. Stark 2013). In particularly frequent collocations such as *je (ne) sais pas*, *ce (n'est pas)*, *ne* deletion seems to be practically generalized. In the case of subject doubling, *ne* retention is unlikely (cf. A.-B. Hansen and I. Malderez (2004: 21) with 9.5% using *ne* for doubled lexical subjects and 3.3% using *ne* for pronoun doubling).

W. Ashby (1981), A.-B. Hansen and I. Malderez (2004), and M.-L. Moreau (1986: 137–160) show a positive correlation between the frequency of the second negative element and *ne* deletion. *Ne* deletion rates are highest with *pas* and decrease with (rather infrequent) *jamais*. This is not a monotonous correlation, as some relatively infrequent items (*personne*, *rien*) have, in some corpora, comparably or even slightly higher rates of *ne* drop compared with *pas*. The same overall tendency is observed in N. Armstrong and A. Smith (2002).

Ne retention is also high with non-finite verb forms (infinitives, participles, gerunds) or imperatives (cf. e.g. W. Ashby 1976: 79%; W. Ashby 1981: 68%; C. Meisner 2010: 80%). See the following examples from our corpus for *ne* with imperatives:

- (8) a. *te stresse pas, l'intercity a aussi 5 de retard [...]* (*ne*-drop)
(‘Don’t worry, the intercity is also 5 minutes delayed’)
- b. *Si ca te di, n’hesit pa ! [...]* (*ne* retention)
(‘If you want to, don’t hesitate’)

With this in mind, two principled hypotheses can be put forward to explain the variation around *ne* in contemporary French. It may be externally triggered (where regional or stylistic factors are the most probable triggers), with some French varieties retaining *ne* and others (e.g. Canadian French or colloquial non-standard French) no longer using it (cf. B. Massot 2010 for a recent support of this “diglossic hypothesis”). Alternatively, it could be subject to language-internal regularities, and its apparent dependency on external factors such as ‘private communication’ is only epiphenomenal. Thus, if clitic pronominal subjects can be shown to be the most important internal triggering factor for *ne* deletion, and if pronominal subjects typically occur in private dialogic contexts (cf. J.W. Du Bois 1987), low rates of *ne* retention can easily be explained without having to assume that the type of communication has a direct influence on *ne* — an assumption that is improbable at any rate (cf. A. Dufter and E. Stark 2002 and 2007 for an extensive discussion).

Assuming that text messages are, in their overwhelming majority, instances of unofficial private communication, that they constitute asynchronous communication and that they are graphic in nature, and given the fact that our corpus was compiled in Switzerland (see next section), we will put forward the following four general hypotheses:

- a. There will be very low *ne*-retention rates, as the corpus reflects Swiss French (cf. B.-B. Fonseca-Greber 2007).
- b. There will be very low *ne*-retention rates, as the corpus reflects mainly unofficial, private communication (cf. N. Armstrong 2002, N. Armstrong and A. Smith 2002, W. Ashby 1981, A. Coveney 2002, C. Meisner 2010).

- c. (Contradictory hypothesis:) There will be slightly higher *ne* retention rates than in phonic corpora of unofficial, colloquial French, in keeping with the regularities observed by R. van Compernelle (2008, 2010) for graphic French in online chats.
- d. Observed language-internal regularities for *ne* deletion and *ne* retention will hold to the same extent in our corpus as in other corpora of contemporary French (cf. R. van Compernelle 2008, 2010). This means that identified morphosyntactic *ne*-deletion triggers like clitic subjects will be relevant in our corpus, as previously shown in other types of private communication in French (e.g. conversations among friends, chats etc.).

2. Data and Methodology

Our study (for quantitative analysis, cf. L. Gubler 2012) is based on the French part of the newly established corpus *SMS4science.ch*. In the autumn/winter of 2009 and in spring 2011, 25,947 authentic text messages were collected with the support of Swisscom (the predominant telecommunication provider in Switzerland), as part of a larger international project (cf. www.sms4science.org). 2,784 participants sent their text messages to a free number, which have been subsequently anonymised and currently form a corpus which has also been made into an online resource for scientific needs (cf. C. Dürscheid and E. Stark 2011 for further details on the corpus). Of the 2,784 participants, 1,316 (*i.e.* about 50%) filled in an anonymous online questionnaire containing socio-demographic information about their sex, age, regional origin, mother tongue, education, reading and writing habits etc. (cf. www.sms4science.ch for information about the average sex, age etc. of the participants). The corpus *SMS4science.ch* is a multilingual corpus. The following table gives an overview of the most important languages represented therein:

(9) Number of text messages (and approximate number of words) per language/variety

	SMSes	Words (approximately)
Swiss German	10,718	218,000
Standard German	7,224	140,000
Standard French	4,628	94,000
Standard Italian	1,474	30,500
Romansh (all varieties)	1,085	25,000

In the 4,628 text messages written in Standard French analysed for this study, we found 1,330 negated verb forms in sentences with the non-clitic negative elements *pas*, *jamais*, *plus*, *rien* and *personne*. We conducted a search on regular expressions (e.g. for <pas> : \bp{1,20}a{0,20}o{0,20}s{0,20}\b), in order to find all graphic variants of these elements. From the results, we discarded all instances of homonymy (e.g. between *une personne*, ‘a person’, and *ne...personne*, ‘nobody’) and all collocations where *ne* could not show up (e.g. *pas de problème*, ‘no problem’, *pas mal*, ‘pretty good’, *pourquoi pas*, ‘why not’, *de rien*, ‘welcome’, *si jamais*, ‘if ever’ etc.):

- (10) a. [...] *merci de cœur, pas selement pr l'argt tu sais* [...] (‘Thank you so much, not only for the money you know’)
 b. *Coucou je t'attends dehors, pas dans le wine bar* [...] (‘Hello, I’ll wait outside, not in the winebar’)
 c. *Et si jamais tu veux déjà venir dormir à Lausanne* [...] (‘And if ever you wanted to sleep at Lausanne’)

We only analysed the three most frequently discussed internal triggering factors for *ne* deletion and retention, as mentioned in Section 1 under (5), *i.e.* subject type (including subject doubling), type and frequency of the second negative element and non-finite verb forms and imperatives, *i.e.* verb forms with no possible overt subject (see examples (6) to (8) above).³ Additionally, we took into consideration deleted subjects (usually expletive ones) with finite verb forms other than the imperative, *i.e.* cases of subject drop, frequently found in CMC:

- (11) *Demain ø faut pas que je boive* [...] (*ne*-drop after missing *il* in *il faut*) (‘Tomorrow I am not allowed to drink’)

3. Results

3.1 Subject type

Among our 1,330 negated verb forms, 312 use *ne*, *i.e.* 23.5%, a figure higher than those observed in comparable phonic corpora of contemporary European French, and also higher than in R. van Compernelle’s (2008) study of European French chat (*ne* retention rate of 16.1%). 1,145 negated verb forms have an overt subject (the others are cases of subject drop, non-finite verb forms or imperatives). The distribution of lexical *vs* pronominal subjects of negated verb forms with or without *ne* is shown in (12) below (*personne* had to be excluded from this analysis, as

3. We did not consider frequent collocations such as *je (ne) sais pas*, *ce (n’) est pas* etc, cf. Moreau 1986, despite the fact that they too are mentioned as crucial for *ne* deletion in the literature.

it figures in 5 out of 6 occurrences in the subject position itself. This explains the total of only 1,139):

(12) Subject type

	<i>-ne</i>	<i>+ne</i>	Σ
Lexical subject (including proper nouns)	27 (34.2%)	52 (65.8%)	79 (100%)
Pronominal subject	850 (80.2%)	210 (19.8%)	1060 (100%)

(12) confirms that the overall tendency is for lexical subjects to favour *ne* retention, similarly to equivalent phonic corpora of French (e.g. 62.3% in A. Dufter and E. Stark 1997 and 56.4% in A.-B. Hansen and I. Malderez 2004). Our findings are comparable to those of R. van Compernelle 2008, however, the R. van Compernelle study shows slightly lower *ne* retention rates both for lexical subjects (59.6%) and pronominal ones (12%). The following examples illustrate this trend, with (13) a, b and c showing the expected *ne* retention (after a lexical subject) and *ne* deletion (after *je* and *ce*), while (13) d is representative of the minority of examples where *ne* is retained even after *je* :

- (13) a. *Alors ma tenue ne t'a pas séduit hier?*
(‘So you didn’t like my outfit yesterday?’)
- b. *Bah, j’ai pas encore trouvé un congrès où je me sentais vraiment à ma place [...]*
(‘Well, I haven’t yet found a conference where I felt really comfortable’)
- c. [...] *Tu fais quoi d bo si c’est pas secret ? [...]*
(‘What are you doing if it is not a secret?’)
- d. [...] *Je ne sais ps si tu es au courant ms ns a demandé [...]*
(‘I do not know whether you know that but [missing subject] asked us’)

After this broad overview, a more precise definition was sought, following C. Meisner (2010). Based on prior corpus studies of *ne* retention in contemporary non-standard French, C. Meisner (2010) subdivides pronominal subjects further into clitic ones (type A), atypical clitics (type B, such as *nous*, *vous*, *elle*, and *elles*, which can also be stressed and serve as an isolated answers to questions), and non-clitic pronouns (type C, such as *quelqu’un*, *qui*, *ça* etc.; lexical subjects constitute her type D). (14) shows the co-occurrences of these four subject types with *ne* in our corpus:

- (14) Subject type: *ne* realization in 4 types according to C. Meisner (2010) in our corpus

Type	Example	– <i>ne</i>	%	+ <i>ne</i>	%	Σ
A: clitic subjects	<i>je, tu, il, ils, ce, (on)</i>	760	80.5	184	19.5	944
B: atypical clitic subjects	<i>elle, nous, vous, elles</i>	21	65.6	11	34.4	32
C: non-clitic subjects	<i>qui, ça, cela</i>	69	82.1	15	17.9	84
D: lexical subjects	<i>mon voisin</i>	27	34.2	52	65.8	79

What is striking here is the relatively high rate of *ne* retention with type B, *i.e.* with *nous, vous, elle* and *elles* (34.4%), based on quite a few examples (32). They can thus be said to constitute a third type of subject with respect to *ne* retention, between lexical subjects and clitic subjects or non-clitic pronominal subjects. Consider the following example of *ne* retention after *nous* from our corpus:

- (15) [...] *nous sommes allés y manger avec Rizvana et nous n'avons pas pu résister à prendre les ravioli Caren* [...]
 ('We went out to eat with Rizvana and could not resist the ravioli Caren')

It remains to be explained why this should be so; there is also the issue of their relatively low frequency compared with the other subject types. Their prosodic characteristics in the group of personal pronouns certainly play a role: they are bisegmental in any French variety and any phonological context, contrary to *e.g.* <je>, [ʒə], which is often reduced to the initial fricative, or <il/ils>, [il], which are often pronounced as simple [i] (cf. C. Meisner 2012). It is important to note for our purposes the striking parallel with C. Meisner's (2010) findings for type A and B, and a similar tendency as in C. Meisner for type D:

- (16) Subject type and *ne* realization in C. Meisner (2010: 1949)

Type	Example	– <i>ne</i>	%	+ <i>ne</i>	%	Σ
A: clitic subjects	<i>je, tu, il, ils, ce, (on)</i>	1421	84	268	16	1689
B: atypical clitic subjects	<i>elle, nous, vous, elles</i>	125	65	66	35	191
C: non-clitic subjects	<i>qui, ça, cela</i>	145	55	121	45	266
D: lexical subjects	<i>mon voisin</i>	26	13	167	87	193

Finally, (17) illustrates the *ne* retention rates for each morphological exponent of overt pronominal subjects with negated verb forms (compared to all lexical subjects) in our corpus:

(17) Subject type: Lexical vs pronominal subjects: morphological exponents

Type	Exponent	– <i>ne</i>	%	+ <i>ne</i>	%	Σ
D	Lexical subjects	27	34.2	52	65.8	79
A	<i>je</i>	501	81.7	112	18.3	613
A	<i>tu</i>	91	79.8	23	20.2	114
A	<i>il</i>	33	64.7	18	35.3	51
A	<i>ce</i>	72	90	8	10	80
A	<i>on</i>	59	75.6	19	24.4	78
A	<i>ils</i>	4	50	4	50	8
B	<i>elle</i>	12	66.7	6	33.3	18
B	<i>nous</i>	1	25	3	75	4
B	<i>vous</i>	8	80	2	20	10
C	<i>ça/cela</i>	62	82.7	13	17.3	75
C	<i>qui</i>	7	77.8	2	22.2	9
	Σ pronominal subjects	850	80.2	210	19.8	1060

By far, the most frequent subject exponent is *je*_{1.SG}, followed by *tu*_{2.SG} and *ce*, *ça/cela*_{3.SG} ('this'). These four elements also have particularly high rates of *ne* deletion (81.7%, 79.8%, 90%, 82.7%). The next highest percentage of *ne* deletion occurs with *vous*_{2.PL} (80%). However, such low absolute numbers (only 10 occurrences in total) mean that this result cannot be reliably seen as representative (the same holds for the relative pronoun *qui*). *On*_{3.SG} ('one', 'we'), following *ça/cela* in absolute numbers, shows only 75.6% of *ne* deletion, a result which cannot be compared to many others. The presence or absence of *ne* after the nasal vowel or the pronounced nasal consonant [-n] of *on* in prevocalic contexts is impossible to detect in phonic corpora, so that these contexts are not usually taken into account. In this respect, a graphic corpus has undeniable advantages. See the following examples of *on* (the first one without, the second one with realized *ne*) taken from our corpus:

- (18) a. *On peut plus compter sur les males jte ture :D*
 ('One cannot count on males any longer I swear :D')
 b. *Decidément on n'est jamais là en même temps.*
 ('Indeed, we are never there at the same time.')

Overall, our data confirm the observed tendency that frequent elements favour *ne* deletion (with the exception of *vous* and *qui*), but we cannot decide whether this is epiphenomenal or a fitting explanation.

We also examined 66 cases of subject drop, which can be seen in (19):

(19) Subject drop

<i>-ne</i>	<i>+ne</i>	Σ
65 (98.5%)	1 (1.5%)	66 (100%)

Although based on a small amount of examples, this seems to confirm the observation of Ashby (1981) according to which dropped subjects almost categorically hinder *ne* realization. The only example in our corpus (shown in (20) below) where a seemingly dropped subject is related to *ne* occurs in a case of conjunction reduction, where the deleted subject of the second conjunct is *Mattis*, a proper name, which explains the presence of *ne*:

- (20) *Mattis a été brave. Mais ne voulait pas dormir*
 ('Mattis has behaved well, but did not want to sleep')

Subject doubling is another important factor influencing *ne* deletion. However, we only found eight cases of subject doubling, which reduces the empirical base of our observation considerably. Nonetheless, these always appear with *pas*, and never with *ne*, cf. (21):

(21) Subject doubling

	<i>-ne</i>	<i>+ne</i>	Σ
Lexical subjects)	4 (100%)	– (0%)	4 (100%)
Pronominal subjects	4 (100%)	– (0%)	4 (100%)

Consider the following examples which demonstrate consistent *ne*-drop, first with a lexical subject and second with a pronominal subject, each doubled by the respective clitics:

- (22) a. *Ma pr vendredi ma mère elle vx pa dc ca ser a ri 2 te libéré [...]*
 ('But on Friday, my mother doesn't want to, so it is not worth making yourself free')
- b. *Demande au autre mé Moi jé pa envi d payé 20fr [...]*
 ('Ask the others but I don't feel like paying 20fr')

These findings could corroborate B. Massot's (2010) hypothesis which states that *ne* is simply incompatible with subject doubling in French. They also confirm the striking rarity of doubling structures in French text messages (cf. E. Stark 2013), as opposed to phonic French. This tendency might indicate that subject doubling and maybe other forms of dislocation and hanging topics (cf. E. Stark 1997, *contra* P. Koch and W. Oesterreicher 2011) are not indicators of low registers in French, but are bound to phonic communication.

Overall, our findings on the influence of subject type on *ne* realization in French text messages confirm the results of the literature on *ne* deletion in contemporary French. Foremost of these is the tendency observed by R. van Compernelle (2008, 2010) that *ne* retention rates are slightly higher in *graphic* discourse than in comparable *phonic* corpora, all other things being equal.

3.2 Form and frequency of the postverbal negative element

Some researchers state that *pas* is particularly prone to *ne* deletion, whereas the non-negative restrictive *que* (*je n'a que deux frères*, 'I only have two brothers') favours its retention, and other postverbal negative elements are situated in between these two extremes (see (5) in Section 1 above, and M.-L. Moreau 1986). R. van Compernelle (2008) listed 14.9% *ne* retention for *pas* and 20.6% for other negative elements. We did not take *que* into account as it is not an expression of sentential negation. We have compared the other postverbal negative elements found in our corpus and obtained the following results:

(23) Postverbal element

	[– ne]	%	[+ ne]	%	Σ
<i>Personne</i>	5	83.3	1	16.7	6
<i>Pas</i>	936	76.8	283	23.2	1219
<i>Plus</i>	25	75.8	8	24.2	33
<i>Jamais</i>	12	75	4	25	16
<i>Rien</i>	40	71.4	16	28.6	56

Surprisingly, the rates of *ne* deletion or retention are quite similar for our five postverbal negative elements; however, this result has to be interpreted cautiously, as only *pas* occurs with sufficient frequency to permit statistically valid statements. For *plus* and *rien*, we cannot state a huge difference in *ne* retention, and overall it seems that in our corpus, form and frequency of the second negative element is not a decisive factor for *ne* retention or deletion.

3.3 Non-finite verb forms / imperative vs indicative verb forms⁴

W. Ashby (1976, 1981) suggests that there is a significant difference between non-finite verb forms and imperatives on the one hand and verb forms with an obligatory subject on the other, especially indicative verb forms. This hypothesis was tested by C. Meisner (2010) and is also the inspiration of our most recent empiri-

4. There are no occurrences of negated *subjonctifs* or conditional verb forms in our corpus.

cal investigation of the French text messages in the Swiss corpus. We separated non-finite verb forms (infinitive, gerunds, participles — 42 altogether) from imperatives, and compared these two groups with verb forms which must have an explicit subject constituent.

The following examples illustrate in a first step the role of non-finite verb forms in *ne* deletion or retention ((24) b and c repeat examples (7) a and b above) :

- (24) a. [...] *Ms ct un plaisir pr ns, désolé de pas etre resté plus longtemps* [...] ('But it's a pleasure for us, sorry for not having stayed longer')
 b. *J'aurai bien voulu resté mais je préfère ne pa etre bien à la maison !!* ('I would have loved to stay but I prefer to feel bad at home!!')
 c. *N'arrivant pas à vous joindre et n'ayant pas de nouvelles* [...] ('As I cannot contact you and do not have any news from you')
 d. *Désolé pas eu le temps de te répondre hier* [...] ('Sorry, I did not have time to answer you yesterday')

(24) a and b show *ne* drop and *ne* retention respectively with the infinitival auxiliary *être* ('to be'); (24) c is a construction formed with the present participle, and (24) d is a case of *ne* drop, but also of subject drop preceding the past participle of *avoir* ('to have'). In the last two examples, we find additional triggers for *ne* retention (very formal style) or *ne* deletion (missing overt subject), which might have influenced the results.

In (25) we can see the overall results for non-finite verb forms:

(25) Non-finite verb forms

	<i>-ne</i>	%	<i>+ne</i>	%	Σ
Finite	988	76.8	300	23.2	1288
Non finite	29	69	13	31	42

The difference in *ne* retention is not very significant (23.3% for finite vs 31% for non-finite verb forms), contrary to what has been observed in W. Ashby (1976, 1981). W. Ashby did not, however, take present participles into account, and he conflated imperatives with non-finite verb forms, which might explain the difference in results. There is a difference in our data between present participles on the one hand (no *ne* deletion; this is confirmed by N. Armstrong and A. Smith 2002, for example), past participles (always *ne* deletion), and infinitives (57.8% of *ne* deletion). However, the absolute numbers for present participles (3) are too low to draw further conclusions from these figures, cf. (26), and the overall trend is not clear enough to consider non-finite verb forms a factor in *ne* retention:

(26) Different non-finite verb forms

	- ne	%	+ ne	%
Infinitive	11	57.8	8	42.2
Pres. participle	0	0	3	100
Past participle	20	100	0	0

Imperatives are a strong trigger for *ne* retention according to W. Ashby (1981) and N. Armstrong and A. Smith (2002). See our results in (27) below (we found 77 negated imperatives in our corpus):

(27) Imperative vs indicative

	- ne	%	+ ne	%	Σ
Indicative	947	78.2	264	21.8	1211
Imperative	41	53.2	36	46.8	77

The following examples illustrate *ne* realization with imperatives in our corpus ((28) a and b repeat examples (8) a and b for the sake of convenience):

- (28) a. *te stresse pas, l’intercity a aussi 5 de retard [...] (ne-drop)*
 (‘Don’t worry, the intercity is also delayed by 5 minutes’)
- b. *Si ca te di, n’hesit pa ! [...] (ne retention)*
 (‘If you want to, don’t hesitate’)
- c. [...] *boire un verre avec gion. Ne m’attends pas. A toute, bisou*
 (‘to have a drink with Gion. Don’t wait for me. See you soon, kiss’)

Interestingly, (28) b and c, which both realize *ne* by using the imperative, differ considerably as to the degree of deviance from standard French orthography. While (28) b consistently drops French silent letters, especially in verb inflections (<di> instead of <dit>, ‘says’, <hesit> instead of <hésite>, ‘hesitate’) and does not show any diacritics at all (<ca> instead of <ça>, ‘that/it’, <hesit> instead of <hésite>, ‘hesitate’), (28) c conforms to a completely standard orthography (except for the use of lower case for the proper name *Gion*). These and similar examples show no correlation between the degree of graphic deviance and the preference for certain patterns of morphosyntactic marking, a fact that we have already considered in other studies on our data (cf. E. Stark 2013, E. Stark and I. Riedel in print).

All in all, in our data, imperatives do not particularly favour *ne* retention; in fact, they seem to be quite irrelevant. Indicative verb forms occur overwhelmingly without *ne*. However, the *ne* retention rate with imperatives is more than double than the *ne* retention rate with indicatives. This result differs from those of W. Ashby (1981) and N. Armstrong and A. Smith (2002), but is comparable to R. van

Compernelle's (2008), who states 16.3% of *ne* retention for indicatives and 19.2% for imperatives. The only point in common concerns the rate of *ne* retention with indicatives, which parallels the absolute *ne* retention rate, both in our data (21.8% vs 23.5%) and in R. van Compernelles's (2008; 16.3% vs 16.1%).

4. Discussion

At the end of Section 1, we put forward four hypotheses, formulated against the background of research on *ne* deletion in contemporary French, both in phonic and graphic corpora. We repeat them here for convenience:

- a. There will be very low *ne*-retention rates, as the corpus reflects Swiss French (cf. B.-B. Fonseca-Greber 2007).
- b. There will be very low *ne*-retention rates, as the corpus reflects mainly unofficial, private communication (cf. N. Armstrong 2002, N. Armstrong and A. Smith 2002, W. Ashby 1981, A. Coveney ²2002, C. Meisner 2010).
- c. (Contradictory hypothesis:) There will be slightly higher *ne* retention rates than in phonic corpora of unofficial, colloquial French, in keeping with the regularities observed by R. van Compernelle (2008, 2010) for graphic French in chats.
- d. Observed internal regularities for *ne* deletion and *ne* retention will hold to the same extent in our corpus as in other corpora of contemporary French (cf. R. van Compernelle 2008, 2010). This means that identified morphosyntactic *ne*-deletion triggers like clitic subjects will be relevant in our corpus, as previously shown in other types of private communication in French (e.g. conversations among friends, chats etc.).

Hypothesis (a) is not corroborated, as the *ne* retention rate in our corpus is relatively high (23.5%) compared with phonic corpora of spoken French, especially Swiss French (only 2.5% in B.-B. Fonseca-Greber 2007), but also compared to R. van Compernelle's (2008) findings on European French in CMC (chats, 16.1%). This may be due to the rather standard-oriented, conservative character of our data, confirmed also in other quantitative studies on subject-verb agreement (cf. E. Stark 2011) or object-past participle agreement (cf. E. Stark and I. Riedel, in print). What we cannot see at all, however, is a regional differentiation, if we compare our data with those from van Compernelle (2008) for European French or R. van Compernelle (2010) for Canadian French in CMC.

Hypothesis (b) is not corroborated either, maybe due to the graphic nature of our rather conservative data which reflects the written standard to a considerable extent.

Hypothesis (c) is corroborated by our data, to a greater extent than in R. van Compernelle's (2008, 2010) findings. This fact demonstrates the necessity of taking into account the medium (phonic *vs* graphic) and not merely the conception (roughly: informal *vs* formal) of new French communication forms (*contra* P. Koch and W. Oesterreicher 2011) when performing variationist research (see below).

Hypothesis (d), which is corroborated by our results for the factor 'subject type', leads us towards a greater understanding of the regularities of *ne* deletion in contemporary French (see below). Form and frequency of the non-clitic second negative element (Section 3.2) are, however, not a triggering factor for *ne* deletion or *ne* retention in our data. This contradicts what has been found in the literature, except for R. van Compernelle (2008), but is potentially due to relatively low numbers for any element except *pas*. Note that the findings, e.g. in W. Ashby (1981), A.-B. Hansen and I. Malderez (2004), and M.-L. Moreau (1986), are also partially contradictory to each other when examined in detail. When it comes to the role of verb forms with no overt subject like non-finite forms and imperatives (see Section 3.3), infinitives and past participles (cf. (24) a and d above) seem to favour *ne* deletion, contrary to present participles, which may be an epiphenomenon of style shift in these constructions. This result, however, cannot easily be compared to earlier findings: W. Ashby (1976, 1981) did not take into account present participles, nor distinguished between non-finite and imperative verb forms. Furthermore, our findings are based on quite low absolute numbers. Imperatives show lower *ne* retention rates than indicatives in our data (cf. 46.8% *vs* 21.8 for indicatives), but are still more frequently found without *ne* than with *ne*. This provides a contradiction to W. Ashby's (1981) and N. Armstrong and A. Smith's (2002) findings, but confirms a trend observed by R. van Compernelle (2008), that imperatives have (slightly) less *ne* deletion than indicatives.

Once we have a closer look at these results alongside the discussion about the language used in CMC and the variable *ne* in contemporary French, it is possible to identify two quite far-reaching implications.

First, the graphic nature of our data has to be taken into account in several respects, as has become repeatedly clear when discussing the results and our four hypotheses. The slightly higher *ne* retention rates that we see in our data when compared with phonic corpora or R. van Compernelle's work on French online chats (2008, 2010) could point to a strong influence from standard orthography in Swiss text messages — which would make our data dubious or not representative of empirical studies of grammatical regularities in contemporary French. However, this rather pessimistic conclusion seems overstated, as the discussion of hypothesis (d) did confirm that language-internal tendencies for *ne* deletion and retention existed in our data, only to a lesser extent, due to the influence of

writing.⁵ *Ne* deletion, like dislocation structures, seems to be partially dependent on the phonic vs graphic nature of the data, as shown by our confirmation of hypothesis c) above. This is in keeping with the slightly higher rates of *ne* retention in R. van Compernelle for French chats compared with phonic communication, and also with the slightly lower *ne* retention rates in the case of imperatives. Our results are in line with R. van Compernelle (2008), but not with findings based on phonic material (W. Ashby 1981, N. Armstrong and A. Smith 2002). If our results can be confirmed by other studies based on a larger amount of non-standard graphic data, this would be an important finding for diachronic research into the issue at hand, as it is by default based on graphic material. It is important to be careful in generalizing its results in order to establish an overall chronology of *ne* deletion in French for example, even when based on non-standard texts. Yet, we cannot take the graphic representation of linguistic information in text messages as an 'authenticity indicator'. Nor can we simply separate on the one hand quite standard-oriented text messages and on the other 'deviant' or 'creative' spelling strategies, and take only the former or only the latter as representative e.g. for Standard or non-standard French varieties, respectively. This becomes clear in many of the examples given above, and especially in those where we find no breach of the French standard orthography, but which integrate features of non-standard French (like *ne* deletion or subject drop, cf. examples (6) c or e and (11)). The same is true of examples (cf. e.g. (7) a, (8 b), (10 a)) which show that deviant spellings and non-standard grammatical features do not necessarily go hand in hand, a fact often neglected in the literature on (French in) CMC. The medium is only one aspect of the 'spoken-written opposition' in variationist research (see the figure under (2) above), *i.e.* writing — or more precisely spelling — has two aspects, a normative and a systematic, grammatically relevant one (writing in unofficial contexts, reflecting actual language use), which have to be discussed separately. Overall, the graphic nature of CMC data in French plays a strong role in research into morphosyntactic variation in French.

Second, the status of *ne* as a sociolinguistic (*i.e.* externally driven) variable is partially called into question by both our own results and other comparable results. Despite the fact that *ne* deletion is less frequent in our data than in comparable phonic corpora and in R. van Compernelle (2008), the main and most robust trigger for it remains the same as has been observed in the literature up to now. Section 3.1 shows that *ne* deletion is much less frequent with lexical subjects (34.2%) than with pronominal ones (80.2%), and that among these, clitical subjects like *je*, *tu* and *ce* favour *ne* deletion even more (80.5%). Our findings, though partially based on very

5. Please note that the majority of our participants have a secondary school diploma or even an academic degree.

few examples, seem to parallel those by C. Meisner (2010) for clitic and 'atypical' clitic subjects. They differ for non-clitic subjects and lexical subjects, but are still in line with them for the overall trends. As C. Meisner (2010) bases her overall findings on phonic data from different situations (casual conversations among friends, classroom interaction, oral exams), this means that the observed regularities are perhaps a strong internal morphophonological regularity of French, *independent of register, style or medium* (cf. C. Meisner 2012). This is evident for subject-doubling (cf. examples in (22) above), which is, in our few examples, strictly incompatible with *ne*, and which seems to be categorically impossible in other types of data (cf. B. Massot 2010). Medium comes into play where the slightly higher *ne* retention rates may demonstrate the influence of orthographic norms in French (see above), but we cannot state any correlation between *ne* drop or *ne* retention and more or less deviant spelling, which is an indirect indicator of casual communication between close friends or family members. A thorough discourse-analytic analysis is now needed to show whether *ne* retention or deletion really has no correlation either with certain forms of address indicating communicative proximity such as *tu* instead of *vous* (second person singular, formal), nor with nicknames, casual greeting formula etc. It is a hypothesis which we would like to see strongly supported by our results.

5. Conclusion

In this paper we have investigated negation marking in a newly established corpus of Swiss text messages in French. In 4,628 original text messages gathered in 2009, we found 1,330 negated verb forms, which were analysed according to the literature on the variable of clitic *ne* drop in contemporary French, and focusing on three frequently mentioned language-internal triggers for *ne* deletion, *i.e.* subject type, form and frequency of the second non-clitic negative element and non-finite / imperative *vs* indicative verb forms. We found an overall *ne* retention rate of 23.5%, somewhat higher than comparable phonic corpora, and also higher than R. van Compernelle's (2008) analysis of European French in chats. We could not confirm a regional or stylistic tendency with the high *ne* drop in our data, which findings from comparable phonic corpora have suggested (especially B.-B. Fonseca-Greber, 2007). Additionally, we took the general regional (Swiss French) and stylistic (unofficial communication) characteristics of the whole corpus into account. The only internal factor that proved relevant was the subject type, more precisely clitic subjects, which caused significantly high rates of *ne* drop (over 80%) in our corpus, just as in other corpora (see C. Meisner, 2010 and in prep.) of contemporary French, irrespective of the medium (graphic *vs* phonic). Moreover, *ne* drop seems to be categorical with subject drop and subject doubling

(cf. Massot, 2010). However, this is based on only a few examples in our data. Other more inconclusive findings like that concerning form and frequency of the second negative element or non-finite / imperative *vs* indicative verb forms seem to parallel those of R. van Compernelle's (2008) on European French used in on-line chats, which might be an indicator of the relevance of the medium for these factors (phonic *vs* graphic data).

More research has to be done now in order to understand, on the one hand, the regularities between *ne* drop and clitic subjects, and, on the other, the specific conditions of the use of French in text messages. Regional variation should also be considered, which will be possible inside the network *sms4science*, thanks to the ongoing setup of French corpora of text messages in Belgium, La Réunion, France (Poitiers, Montpellier, Lyon) and Québec. Finally, a sociolinguistic study is needed which seeks to correlate our results with the language-external socio-demographic data available to us (age, sex, education etc.), in order to learn more about the potential character of the negation marker *ne* as a sociolinguistic variable.

References

- Anis, Jacques. 2007. Neography. Unconventional Spelling in French SMS Text Messages. In *The Multilingual Internet. Language, Culture and Communication Online*, B. Danet, S. Herring (eds.), 87–115, New York: Oxford University Press.
- Armstrong, Nigel. 2002. Variable deletion of French *ne*. A cross-stylistic perspective. *Language Sciences* 24: 153–173.
- Armstrong, Nigel; Alan Smith. 2002. The influence of linguistic and social factors on the recent decline of French *ne*. *Journal of French Language Studies* 12: 23–41.
- Ashby, William. 1976. The loss of the negative morpheme NE in Parisian French. *Language* 57: 647–687.
- Ashby, William. 1981. The loss of the negative particle *ne* in French: A syntactic change in Progress. *Language* 75: 674–687.
- Ashby, William. 2001. Un nouveau regard sur la chute du *ne* en français parlé tourangeau. S'agit-il d'un changement en cours? *Journal of French Language Studies* 11: 1–22.
- Auger, Julie. 1994. *Pronominal Clitics in Québec Colloquial French: A Morphological Analysis*. Pennsylvania. Unpublished.
- Coveney, Aiden. 2002. *Variability in Spoken French. A Sociolinguistic Study of Interrogation and Negation*. Bristol UK/Portland USA: Elm Bank.
- Cresti, Emanuela; Massimo Moneglia. 2005. *C-Oral-Rom. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.
- Culbertson, Jennifer. 2010. Convergent evidence for categorical change in French: From subject clitic to agreement marker. *Language* 86(1): 85–132.
- Diller, Anne-Marie. 1983. Subject NP Structure and Variable Constraints: The Case of French NE Deletion. In *Variation in the Form and the Use of Language*, R. Fasold (ed.), 167–174, Washington: Georgetown University Press.
- Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63(4): 805–855.

- Dürscheid, Christa. 2002. E-Mail und SMS — ein Vergleich. In *Kommunikationsform E-Mail*, A. Ziegler; C. Dürscheid (eds.), 93–114, Tübingen: Stauffenburg.
- Dürscheid, Christa; Elisabeth Stark. 2011. SMS4science: An International Corpus-based Texting Project and the Specific Challenges for Multilingual Switzerland. In *Digital Discourse. Language in the New Media*, C. Thurlow, K. Mroczek (eds.), 299–320, Oxford: Oxford University Press.
- Dufter, Andreas. 2012. Zur Geschichte de *ne*-Absenz in der neufranzösischen Satznegation. In: *L'absence au niveau syntagmatique. Fallstudien zum Französischen*, L. Fesenmeier, A. Grutschus, C. Patzelt (eds.), 131–158. Frankfurt am Main: Klostermann.
- Dufter, Andreas; Elisabeth Stark. 2002. La variété des variétés. Combien de dimensions pour la description? Quelques réflexions à partir du français. *Romanistisches Jahrbuch* 53: 81–108.
- Dufter, Andreas; Elisabeth Stark. 2007. La linguistique variationnelle et les changements linguistiques 'mal compris'. Le cas de la 'disparition' du *ne* de négation. In *Etudes sur le changement linguistique en français*, C. Marchello-Nizia, B. Combettes (eds.), 115–128, Nancy: Presses Universitaires de Nancy.
- Ernst, Gerhard. 1985. *Gesprochenes Französisch zu Beginn des 17. Jahrhunderts. Direkte Rede in Jean Héroards "Histoire particulière de Louis XIII" (1605–1610)*. Tübingen: Niemeyer.
- Fonseca-Greber, Bonnibeth Beale. 2007. The emergence of emphatic *ne* in conversational Swiss French. *Journal of French Language Studies* 17(3): 249–275.
- Grevisse, Maurice; André Goosse. 1993. *Le bon usage: Grammaire française*. Paris: Duculot.
- Gubler, Lisa. 2012. *L'omission du ne de négation dans les SMS suisses francophones*. Unpublished Master thesis, University Zurich.
- Hansen, Anita Berit; Isabelle Malderez. 2004. Le *ne* de négation en région parisienne. Une étude en temps réel. *Langage et société* 107: 5–30.
- Haspelmath, Martin. 1997. *Indefinite Pronouns*. Oxford: Oxford University Press.
- Koch, Peter; Wulf Oesterreicher. 2001. Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit. In *Lexikon der Romanistischen Linguistik*, G. Holtus, M. Metzeltin, C. Schmitt (eds.), Vol. 1.2., 584–637, Tübingen: Niemeyer.
- Koch, Peter; Wulf Oesterreicher. 2011. *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Berlin/New York: de Gruyter.
- Larivée, Pierre. 2004. *L'association négative. Depuis la syntaxe jusqu'à l'interprétation*. Genève: Droz.
- Martineau, France. 2009. Modeling Change: A Historical Sociolinguistics Perspective on French Negation. In *Corpus Analysis and Variation in Linguistics*, J. Durand, Y. Kawaguchi, M. Minegishi (eds.), 159–178. Amsterdam: John Benjamins.
- Martineau, France; Raymond Mougeon. 2003. A sociolinguistic study of the origins of NE deletion in European and Quebec French. *Language* 79(1): 118–152.
- Massot, Benjamin. 2010. Le patron diglossique de variation grammaticale en français. *Langue Française* 168: 87–106.
- Meisner, Charlotte. 2010. A Corpus Analysis of Intra- and Extralinguistic Factors triggering *ne*-Deletion in Phonic French. In *Congrès Mondial de Linguistique Française — CMLF 2010*, Frank Neveu et al., (eds.), 1943–1962, Paris: Institut de Linguistique Française.
- Meisner, Charlotte. 2012. *Ne? Eine Korpusanalyse zur Variation im Ausdruck der französischen Negation*. Talk given the 3 May 2012 at the Universitaet Zuerich.
- Meisner, Charlotte. In prep. *La variation pluridimensionnelle. Une analyse de l'absence et présence du ne de négation en français contemporain*. PhD thesis, Universitaet Zuerich.
- Moreau, Marie-Louise. 1986. Les séquences préformées: entre les combinaisons libres et les idiomatismes. Le cas de la négation avec ou sans *ne*. *Le Français Moderne* 54: 137–160.

- Pohl, Jacques. 1968. *Ne* dans le français parlé contemporain: les modalités de son abandon. In *Actes du XIe Congrès International de Linguistique et de Philologie Romanes*, Vol. 3., 1346–1353, Madrid.
- Riegel, Martin; Jean Christophe Pellat; René Rioul. 1994. *Grammaire méthodique du français*. Paris: Presses Universitaires de France.
- Stark, Elisabeth. 2011. La morphosyntaxe dans les SMS suisses francophones: Le marquage de l'accord sujet — verbe conjugué. *Linguistik Online* 48. http://www.linguistik-online.de/48_11/stark.htm
- Stark, Elisabeth. 2013. Clitic subjects in French text messages: Does technical change provoke and/or reveal linguistic change?. In *Deixis and Pronouns in Romance Languages*, K. Jeppesen Kragh, J. Lindschouw (eds.), 147–169. Amsterdam/Philadelphia: Benjamins.
- Stark, Elisabeth; Isabelle Riedel. In print. L'accord du participe passé dans les SMS francophones du corpus SMS suisse. Ms.
- Thurlow, Crispin. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* 1–1: 1–27.
- Thurlow, Crispin; Michèle Poff. 2013. Text Messaging. In *Handbook of the Pragmatics of CMC*, Susan C. Herring, D. Stein, T. Virtanen (eds.), Berlin/New York: Mouton de Gruyter, 163–190.
- Van Compernelle, Rémi A. 2008. Morphosyntactic and phonological constraints on negative particle variation in French-language chat discourse. *Language Variation and Change* 20(2): 317–339.
- Van Compernelle, Rémi A. 2010. The (slightly more) productive use of *ne* in Montreal French chat. *Language Sciences* 32: 447–463.
- Walther, Joseph B. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23: 3–43.
- www.franctext.fr
www.sms4science.ch
www.sms4science.org

Summary

This study investigates the drop of the first clitic element, *ne*, of French sentential negation. It is based on 4,628 French text messages taken from the newly established corpus *sms4science.ch*. It shows that regional or stylistic factors do not play a major role in triggering *ne* deletion or *ne* retention, and that the only relevant language-internal factor is subject type, more precisely clitic subjects (also in subject doubling structures) and subject drop, which seem to favour or even trigger categorically *ne* deletion. Our findings are in keeping with those by R. van Compernelle (2008) on European French in online chats, and thereby indicate partially specific regularities in French computer-mediated communication (CMC). The findings question traditional assumptions on *ne* drop as a variety marker in French, irrespective of the phonic or graphic nature of the data (cf. P. Koch and W. Oesterreicher 2011). In fact, even in graphically deviant data, *ne* drop seems to follow the observed robust language-internal regularities. Some data seem to indicate that the graphic nature of SMSes plays a role also in *ne*-drop (cf. similar conclusions by R. van Compernelle 2008 for chats); this has to be checked by examining a larger number of French text messages.

“i didn’t spel that wrong did i. Oops”

Analysis and normalisation of SMS spelling variation

Caroline Tagg*, Alistair Baron**, and Paul Rayson**

*University of Birmingham / **Lancaster University

Introduction

Spelling variation, although present in all varieties of English, is particularly prevalent in SMS text messaging (D. Crystal 2008) just as it is in historical varieties (such as Early Modern English), child and non-native learner language, and computer-mediated communication (such as instant messaging). Researchers argue that the choices made regarding spelling variants in SMSes are functional, principled and meaningful. They are principled in the sense that they follow orthographic principles of English and as such reflect and extend existing patterns of variation across historical and contemporary texts (T. Shortis 2007); and meaningful because they contribute to the performance of social identities (C. Tagg 2012). As yet, however, little attempt has been made to empirically validate SMS spelling patterns (for most languages, with the notable exception of French [e.g. C. Fairon and S. Paumier 2006]) and verify the extent to which they mirror those in other texts.

In this paper, we report on the use of VARD2 (A. Baron and P. Rayson 2009) to manually normalise the spelling variation in CorTxt (C. Tagg 2009), a corpus of over 11,000 SMSes collected in the UK between 2004 and 2007. A second tool, DICER (Discovery and Investigation of Character Edit Rules) (A. Baron *et al.* 2009a), was used to analyse the normalisations made in VARD2. DICER extracts letter replacement rules which can transform the variant form to its standard equivalent and builds a detailed database of these rules and their frequencies. Through examining DICER’s analysis and categorising the spelling variants we can build a pattern of spelling trends in SMS for comparison with other text types. Our longer term aim is not only to better understand the nature of SMS spelling, but to construct a set of spelling rules which can be used to automatically normalise spelling in larger SMS corpora.

The paper is structured as follows. We start in Section 1 with an overview of research into SMS text messaging which highlights the significance of spelling

variation as well as the lack of empirical validation. In Section 2 we give a brief description of VARD2 and DICER, before describing CorTxt in Section 3. The manual normalisation process and some initial findings are discussed in Section 4, with the spelling patterns found with DICER presented in Section 5. Finally, we conclude the paper and look to future research directions.

1. The significance of SMS spelling variation

The popular view of text messaging — or at least that represented in the print and online media — has tended to be that spelling variation is chaotic and likely to hinder meaning rather than create it. C. Thurlow's (2006) survey of newspaper coverage of 'issues related to young people, language and new technology' (p. 671) included 101 articles mainly from the UK and the US. Two of the recurring assertions across the articles were that 'textese' represents a 'decisive and dramatic break with conventional practice' (p. 672) and that these radically distinct practices can be linked to falling standards in literacy. As C. Thurlow (2006: 677) acknowledges, it should also be pointed out that newspapers have since picked up on and reported academic work that suggests a more positive effect on literacy. Take, for example, an article entitled 'Texting boosts children's literacy ;-)' in *The Sunday Times* (May, 2008). Nonetheless, it seems fair to suggest that, to the casual reader, the media remains sceptical to some extent of such reports. To cherry pick a particularly plum example from a British newspaper, in an article from *The Guardian* in 2007 entitled 'I h8 txt msg: how texting is wrecking our language', John Humphrys writes that texters are:

"vandals who are doing to our language what Genghis Khan did to his neighbours eight hundred years ago.

They are destroying it: pillaging our punctuation; savaging our sentences; raping our vocabulary. And they must be stopped."

The sentiment is somewhat extreme but it reflects many of those captured in Thurlow's earlier study. However, as will be described shortly, these popular assumptions — that 'textese' is unique, disorganised and damaging — have been challenged by academic researchers.

As well as assuming the destructive and arbitrary nature of SMS spellings, coverage such as that mentioned above may also feed into popular beliefs as to the most common forms of spelling variation in texting, despite the lack of much evidence that such forms are actually very frequent (e.g. C. Thurlow and A. Brown 2003). In the above examples, we can point to the use of 8 in *h8* and the consonant writing (*txt msg*); as well as the emoticon (;-)) in the more positive report. Even

D. Crystal (2008), who doubts the use of some extreme forms of abbreviation reported in the press (and suggests the benefits of texting for literacy), appears to perpetuate this conception of textese in the title of his popular book, *Txt Msg: the Gr8 Db8*. He also provides a list of reported English spelling variants (pp. 189–198) which include the kind of abbreviated phrases he nonetheless feels are unlikely to occur in ‘real’ text messages (*AFAIK, ASLMH, BION, ICWUM, PTMM, TTYL8R*).

Scholarly research based on (relatively small) datasets makes several observations which challenge the popular perspective on texting. Firstly, as mentioned above, variation in spelling is found to occur far less often than is popularly assumed. C. Thurlow and A. Brown (2003), for example, find that abbreviations occur in less than 20% of the overall message content, and other studies produce comparable figures (N. Doring, 2002; A. Deumert and S. Masinyana, 2008). Where variation in spelling does occur, however, the words most likely to be shortened are short, common ones. As R. Grinter and M. Eldridge (2003) note in their study of teenage texting practices:

“Instead of being long challenging phrases offered by dictionaries, the teenagers recorded shortening simple words such as tomorrow and weekend, which often appeared in messages discussing plans. Other commonly shortened words included school, football, Internet, lessons, and homework.”

The implications are that spelling variation emerges from use (that is, by texters abbreviating words that they commonly use) and that interlocutors are likely to understand the recurring ‘respelt’ versions. In other words, spelling variation is functional, principled and meaningful.

Spelling variants or ‘respellings’ are functional in the sense meant by T. Shortis (2007) in his discussion of ‘Txt’ (the language of texting) in that they emerge during interaction as a response to immediate functional demands. That is, the respellings are neither prescribed or learnt but used and picked up on in response to particular interlocutors in specific contexts. One outcome of this is that different groups will converge in practice and as such develop their own ‘codes’ (Tim Grant, pers comm.).¹ This sensitivity to context means that it is unlikely (although not impossible) that spelling variation would be regularly misunderstood or that it would hinder meaning between participants who are part of the same texting network or community. Integral to this is the argument that, although the tendency towards abbreviation may be shaped in part by the affordances of the technology

1. Dr Tim Grant’s project at the Centre for Forensic Linguistics at Aston University involved participants contributing text messages online. He used a ‘snowballing technique’ whereby participants were recruited through existing participants, thus enabling differences between ‘networks’ to emerge (see <http://www1.aston.ac.uk/lss/subject-areas/english/activities/texting-study/>).

(Y. Hård af Segerstad, 2002), texters who abbreviate are seen to respond also to communicative or 'discursive' demands (C. Thurlow and A. Brown 2003). So, abbreviations may be encouraged not only by the limited character allowance but also by the pressure to reply quickly, to use recognised forms, and perhaps to do so whilst on the move or engaged in other activities. And of course abbreviation is not the only function of spelling variation in texting. C. Thurlow and A. Brown (2003)² posit that 'brevity and speed' is just one motivation among three social 'maxims' which encourage variation in spelling:

- a. 'brevity and speed' (seen in lexical abbreviation including letter-number homophones; and the minimal use of capitalisation, punctuation and spacing);
- b. 'paralinguistic restitution' (such as the use of capitals to indicate emphasis or loudness, or multiple punctuation, which 'seeks to redress the apparent loss of such socio-emotional or prosodic features as stress and intonation')
- c. 'phonological approximation' (that is, attempts to capture informal speech, which 'engenders the kind of playful, informal register appropriate to ... text-messaging').

As C. Thurlow and C. Brown point out, respellings such as <ello>, <goïn>, and <bin> fulfil the need for brevity and speed as well as — in this case — approximating spoken conversation (they are examples of phonological approximation). However, in some cases, the latter two principles will override the concern with abbreviation. For example, with a form such as <nooooo!> the maxim of paralinguistic restitution requires more key presses than a simple <no> (particularly if the form is not recognised by the predictive text dictionary); as does a phonological approximation such as <nope>.³

Spelling variation is principled in the sense that it follows the conventions of the language's orthography, and can only be meaningful to the extent that it does so. For instance, in the case of English, to take M. Sebba's (2007) example, scratching *Down with skool* on to a school desk may be seen as a sign of rebellion (signalling a lack of education, and a disregard for the rules) — whilst scribbling *Down with sguul* would not. <Sguul> does not convey the same meaning as <skool> because it departs too far from English orthographic principles. One implication of this is that, because choices in spelling are constrained in this way, spelling variation practices in texting will reflect those in other written discourses. And, indeed, it appears that parallels can be drawn between Txt spellings and those observed in

2. Thurlow and Brown's article appeared in *Discourse Analysis Online* and as an online journal there are no page numbers.

3. In this paper, we follow the convention of marking orthographic representations with <> (see, e.g., Sebba 2007), so that *you* is a word which in texting is often spelt <u> or <ya>.

other texts. Other domains in which variation in spelling is particularly prevalent include the following:

- Historical varieties such as Early Modern English (M. Görlach 1991; A. Baron *et al.* 2009b).
- Advertising: such as the phonetic use of <k> in brand names such as Kwik Mart, an ongoing practice that can be traced back to the beginning of the twentieth century (L. Pound, 1925; V. Cook 2004); and the homonym <u> for *you* in, for example, ‘High Class Shoe Repairs While U Wait’ and ‘Phones 4U’ (V. Cook 2004).
- Dialogues in novels: eye dialect used in novels to represent a character’s non-standard or regional speech and lower class or uneducated status (R.-M. Weber, 1986; D. Crystal, 2003), such in this exchange between Tess and the ‘dark queen’ in T. Hardy’s *Tess of the d’Urbervilles* (1891):

“How darest th’ laugh at me, hussy!” she cried.

“I couldn’t really help it when tootherers did,” apologized Tess, still tittering.

“Ah, th’st think th’ beest everybody, dostn’t, because th’ beest first favourite with He just now! But stop a bit, my lady, stop a bit! I’m as good as two of such! look here here’s at ‘ee!”

- Pre-electronic written communication: for example, various spelling variants occurred in ordinary informal letters — A. Kessler and A. Bergs’ (2003) study of nineteenth century love letters from working class girls included the following: phonetic spellings such as <bcoz> (*because*) and <luv> (*love*); graphic symbolisations of kisses (xx) and roses, and grammar and spelling mistakes.
- Computer-mediated communication, such as instant messaging (R. Ling and N. Baron 2008).

It also appears that patterns of spelling variation are similar across diverse English-speaking communities. In their study of text messages written in English by Xhosa speakers in South Africa, A. Deumert and S. Masinyana (2008), for example, find respellings that reflect those identified in the US by C. Thurlow and A. Brown (2003), and posit similar motivations. Similar spelling motivations and patterns have also been documented in other languages, such as French (J. Anis, 2007).

Seen from a sociocultural perspective that posits orthography, like literacy, as a social practice (M. Sebba 2007), the fact that users are drawing on a constrained set of choices means that the respelling of words is always meaningful in the sense that it indexes particular identities. Spelling variation in SMSes has been described as meaningful because it contributes to how texters portray themselves through texting, and how they position themselves in relation to the people they text. C. Tagg (2012), for example, suggests that respelling is one resource that texters also

draw on in seemingly stylised ways when performing complex social identities through text messaging. Attempts at ‘paralinguistic restitution’ like *I’m parked next to a MINI!!!* and the ‘phonological approximation’ in *exercise, yeah, i kinda remember wot that is, hmm* suggest a speech-like informality which indexes a casual and intimate relationship with interlocutors (C. Tagg 2012). Also evident in SMS is a sense of play or fun. In C. Tagg’s (2012) study, texters in their twenties and thirties were described as deliberately and cleverly manipulating idioms as well as spelling to convey their identities as playful and highly literate individuals: *I’ve just cooked a rather nice salmon a la you*. Finally, a number of other studies describe a role for spelling variation in indicating group identity and membership, particularly in reference to adolescent groups (E. Kasesniemi and P. Rautiainen, 2002; R. Ling 2005). To return to the South African study cited above, alongside what may be international Txt spellings, A. Deumert and S. Masinyana found evidence of forms which index a local identity. The message below shows how international and local forms combine.

“Brur its 2bed one matras my darling is going 2 put me in shid in church. My money i have save have been decrease due 2 da Aunt Mayoly’s funeral,& miner probst. So da case is coming very soon 3months preg. I’ll c then. Sharp.”

(A. Deumert and S. Masinyana 2008: 126)

Locally-motivated forms (underlined above) include *Brur*, which is Afrikaans for brother; *da* for *the* (in an African-American pronunciation); and *shid*, which reflects the South African pronunciation of *shit* with a voiced stop. As such, spelling variation serves to strengthen social networks and a sense of local identity. Far from damaging language and literacy, studies suggest that ‘Txt’ enriches literacy practices and extends, in T. Shortis’s (2007) words, the ‘orthographic palette’ of spelling options from which texters can draw in performing identity.

In summary, the previous research presented above suggests that analysing the principles and patterns of SMS spelling variation is necessary in understanding how meaning is made and identity conveyed in text messaging. Recent research from the computational linguistics community has focussed on automatic normalisation rather than linguistic analysis of the spelling variation in SMS text. F. Yvon (2010) used finite-state transducers and R. Beaufort et al (2010) combined spell checking and machine translation approaches to normalise the orthography of French SMSes. Little attempt has yet been made, however, to identify and categorise patterns of spelling variation across large datasets. It is this gap that the present study seeks to fill.

2. VARD and DICER

VARD (A. Baron and P. Rayson 2009) and DICER (A. Baron *et al.* 2009a) are both tools originally developed to normalise and analyse spelling variation in Early Modern English corpora (see e.g. A. Lehto *et al.* 2010). However, they have been adapted and trained to deal with other types of spelling variation, such as child language data (A. Baron and A. Rayson 2009) and second language learner data (P. Rayson and A. Baron 2011). For this study, VARD and DICER will be used to study SMS spelling variation found in CorTxt.

VARD⁴ (version 2.3 was used for this study) functions as a pre-processor for other corpus linguistic tools by finding and inserting standard modern equivalents for spelling variants to make corpus analysis more accurate. Studies have shown that spelling variation in historical texts has a negative impact on the accuracy of key words (A. Baron *et al.* 2009b) and key word clusters analysis (M. Palander-Collin and M. Hakala 2011), POS tagging (P. Rayson *et al.* 2007) and semantic annotation (D. Archer *et al.* 2003). It is sensible to assume that spelling variation in other language varieties, such as SMS, is likely to have a similar impact on these corpus analysis techniques. VARD2 can be used to both manually and automatically normalise single texts or a whole corpus. The manual normalisation mode also acts as a training mechanism, with the tool learning how to best use its normalisation techniques to deal with the spelling variation in a given corpus. For each normalisation made, an XML tag is inserted into the saved text with the original form saved as an attribute, for example:

<normalised orig="l8r">later</normalised>

This format allows most corpus linguistic tools to use the normalised form in their analysis, whilst the original spelling is maintained. This pairing of the spelling variant to the chosen normalised form can also be used to provide data to the second tool used in the analysis produced for this study — DICER. The specific character differences between the variant and normalised forms are analysed by DICER and spelling rules are created; e.g. for the above example, the rule “Substitute ‘8’ for ‘ate’” would be created. How often these spelling rules are needed for the spelling normalisations in a corpus, and in which positions⁵ and context within the

4. For further details about how VARD2 functions, see Baron and Rayson (2009) and the software’s website (<http://ucrel.lancs.ac.uk/vard>), where it can also be downloaded free for academic use.

5. A position of *First*, *Second*, *Middle*, *Penultimate* or *End* is recorded.

variants, are collated into a database and presented in a series of webpages⁶ for analysis. The rules created and their frequencies are useful for the investigation of spelling trends (as presented below), but also for finding “letter replacement rules” which can be added to VARD2 to improve its normalisation performance (see A. Baron *et al.*, 2009a).

3. Data and participants

CorTxt, the corpus used in this study, was collected for the purposes of linguistic analysis by C. Tagg (2009). The corpus specifications are shown in Table 1. Friends and family of the researcher submitted messages they had received as well as those they had sent and so the corpus includes texts from an extended ‘network’ which reaches beyond the researcher’s own interlocutors to texters in various places across Britain. At the same time, findings from this extended network must be generalised with care (as with any texting group): the corpus is rapidly dating,

Table 1. CorTxt description

No of messages	11,067
No of words	190,516
Average no of words per message	17.2 words
Collection period	March 2004 — May 2007 (3 years, 2 months)
Collection methods and procedures	From friends and family (10,626 messages) AOL anonymous online public forum (441). A further 441 messages came from an anonymous public forum provided by AOL for forwarding text messages. After it was discontinued, people kept using it: that is, messages used in this study were those which were sent to the service but not forwarded.
Composition of texters	Mainly British English speakers, aged 19–68, professionals and students Male = 41%; F = 59%
Language	English (mainly British English)
Type of communication	Mainly personal communication, although some business text messages evident, including: <i>Hi Are you interested in working taster day at Cornwall College, Camborne 27.4.06 4hrs@£6 per hour? Ring if interested. NAME173</i>

6. See <http://corpora.lancs.ac.uk/dicer>

contributors were older than in many studies, they comprised generally well-educated professionals or students, and with one or two exceptions they had English as their first language. In particular, it is possible that the spelling practices of this relatively literate group may involve a higher degree of sophistication than, for example, in text messages composed by adolescents. As described above, we expect that many of the spelling variants will reflect general patterns; at the same time, some spellings will emerge from the practices of these particular groups and this will have implications for the training of a normalisation tool.

4. Manual normalisation

In order to produce patterns for DICER to analyse, VARD 2.3 was used to manually normalise 2,430 randomly selected messages (41,342 tokens) from CorTxt, equating to around a fifth of the whole corpus. The randomly selected messages were split between the three authors for normalisation, except for 5% of the total messages which were included in all three sample sets in order to calculate inter-rater agreement scores. From this overlap, 200 normalisations were made by at least one person. 75% of these normalisations were fully agreed, in that all three agreed (a.) that a normalisation should be made and (b.) on the word to normalise to. Of the 50 normalisations not fully agreed:

- In 28 instances, only one person decided to make a normalisation.
- In 19 instances, two decided to make a normalisation (and agreed on the word to normalise to), the other person made no normalisation.
- In the 3 other instances, a normalisation has been made by all three people, but two have chosen a different word to normalise to than the other person.

Hence, when at least two people have made a normalisation (172 cases), 98.26% have agreement on which word to normalise to, and when all three have agreed to make a normalisation (153 cases), 98.04% have agreement on the word to normalise to. Therefore, we can conclude that overall there was very high inter-rater agreement on actual normalisations made, with a small amount of disagreement on whether or not some words should be normalised.

In total, the normalisation procedure resulted in 3,166 spelling variant normalisations, meaning just 7.66% of the tokens in the sample were considered spelling variants and normalised with VARD. Furthermore, it was found that around half of the messages (1,217) contained no spelling variants at all. These low quantities of spelling variation confirm previous smaller-scale studies which found that spelling variation occurs far less often than is popularly assumed (N. Doring 2002; C. Thurlow and A. Brown, 2003; A. Deumert and S. Masinyana 2008).

Of the 3,166 spelling variants normalised, 322 (10.17%) could be considered “real-word errors”, *i.e.* the variant spelling matches an unintended dictionary word. This figure could be considered quite low compared to previous studies highlighting real-word error rates; for example, J. Pedler and R. Mitton (2010) found that 31.4% of dyslexic writers’ spelling errors were real-words and P. Rayson and A. Baron (2011) found that 21.9% of second language learner errors were real-words. However, as real-word errors are inherently difficult to detect automatically, this low figure requires further verification.

5. DICER Analysis

DICER was used to analyse the 3,166 spelling variant normalisations produced. The frequencies of the spelling normalisations examined resemble a common Zipfian distribution for word frequencies, with some highly frequent spelling variants found and a long tail of unique occurrences. There are 744 spelling normalisation types, with 525 being unique in our sample.⁷ The 10 most frequent spelling normalisations are given in Table 2.

The DICER analysis provides various results which show the spelling trends found in our sample and highlight how difficult automatic normalisation is for this type of spelling variation. For instance, only 28.5% (42% for types) of spelling variants could be normalised by a single character edit.⁸ This is much lower than similar rates found for common spelling errors, *e.g.* R. Mitton (1987) noted that 69% of spelling errors from a range of sources could be corrected with just a single character edit, while earlier studies noted higher rates still. Indeed, many spelling correction methods rely on this by finding dictionary words which are one edit away from a given spelling error (*e.g.* K. Church and W. Gale 1991). The reason why so many of our spelling variants require more than one character edit may be due to a high number of different forms of shortening. This view is supported by the high number of normalisations which require insertion of characters; 72% (57% for types) of normalisation operations⁹ were insertions (27% were substitutions and

7. A spelling normalisation is considered unique when the combination of spelling variant and normalisation is unique.

8. An edit here is an insertion, deletion or substitution of one character, as used in Levenshtein Distance (Levenshtein, 1966).

9. There is often more than one normalisation operation per spelling variant required, *e.g.* <dn̄t> requires the operations *Insert o* and *Insert ’* to arrive at the normalised form *don’t*. In total there were 3,701 normalisation operations and 930 when each variant-normalisation pair was considered once (*i.e.* for types).

Table 2. Most common spelling variant normalisations

Variant	Normalisation	Frequency
u	you	666 (21.04%)
2	to	187 (5.91%)
r	are	95 (3.00%)
4	for	85 (2.68%)
ur	your	75 (2.37%)
tomo	tomorrow	74 (2.34%)
its	it's	71 (2.24%)
b	be	69 (2.18%)
c	see	52 (1.64%)
im	i'm	43 (1.36%)

less than 1% were deletions), compared to 33% of normalisations requiring insertion in a study by R. Mitton (2008: Table 3).¹⁰ This will be discussed further in our analysis of spelling variation categories.

In terms of specific normalisation rules found by DICER, the most frequently required rules (for variant tokens) largely reflect the most common spelling variants found, as shown in Table 2. The ten most frequent rules are shown in Table 3, with all but *Insert g* and *Insert space* relating directly to the top 10 most common spelling variants (see, for example, *Insert yo*), although most rules also incorporate other variants, particularly *Insert 'e* and *Insert e*. In other words, the normalisation rules appear to be shaped by the very frequent occurrence of particular variants, chiefly <u>, rather than the general application of a particular rule across variants. Considering only each individual spelling variant once, the top 10 most frequent rules in terms of types are given in Table 4. There is quite a lot of overlap between the two top 10 lists, although *Insert space* is the most frequent for types (clearly due to the possibility to omit spaces between potentially any combination of words). *Insert o* and *Insert h* are also present in the types top 10, but not for tokens. As previously mentioned, these rules not only highlight the spelling trends found in our sample, but provide possible rules which could be used to assist in spelling normalisations both manually and automatically. The rules which appear most frequently for tokens will help with normalising the most frequent spelling variants

10. The table given by Mitton shows "Omissions" and "Insertions"; these are in terms of how the misspelling is changed from the correct word, *i.e.* an omission means a letter is missing from the misspelling. The results from DICER are in terms of how the misspelling (or variant) needs to change to successfully correct (or normalise) it.

Table 3. Top ten most frequent normalisation rules in terms of tokens

Rule	Frequency	Top position	Examples
Insert yo	802 (21.67%)	Start (100%)	u, ur, u'd
Insert '	387 (10.46%)	Penultimate (72.61%)	its, dont, thats
Insert e	351 (9.48%)	End (87.46%)	r, b, hav
Substitute 2 → to	249 (6.73%)	Start (99.60%)	2, 2day, 2moro
Insert a	162 (4.38%)	Start (91.98%)	r, n, bout
Insert g	122 (3.30%)	End (98.36%)	goin, comin, mornin
Insert space	109 (2.95%)	Middle (81.65%)	thankyou, aswell, togo
Substitute 4 → for	89 (2.40%)	Start (100%)	4, 4ward, 4got
Insert rrow	74 (2.00%)	End (100%)	tomo
Substitute c → see	54 (1.46%)	Start (100%)	c, cing

Table 4. Top ten most frequent normalisation rules in terms of types

Rule	Frequency	Top position	Examples
Insert space	97 (10.43%)	Middle (83.51%)	ifnot, outthatway, togo
Insert e	62 (6.67%)	End (64.52%)	pls, wher, lv
Insert '	61 (6.56%)	Penultimate (70.49%)	cant, didnt, hes
Insert g	56 (6.02%)	End (96.43%)	mornin, lookin, usin
Substitute 2 → to*	34 (3.66%)	Middle (97.06%)	want2go, need2say, what2do
Insert o	17 (1.83%)	Second/Middle (35.29%)	cud, dnt, hw, anymre
Insert yo	16 (1.72%)	Start (100%)	u, uve, u'll
Substitute 2 → to	16 (1.72%)	Start (93.75%)	2day, 2mora, 2gether
Insert h	16 (1.72%)	Second (43.75%)	wen, wich, wat
Insert a	15 (1.61%)	Start (53.33%)	bout, ne (any), n (and)

* Here the rule is substitute 2 with to surrounded by white space.

found (e.g. those given in Table 2). On the other hand, rules that appear most frequently for types will help with normalising a larger variety of spelling variants.

5.1 Categories of spelling variation

The DICER analysis was also used to assist in categorising variants, the ultimate aim of which would be to create a taxonomy of SMS spelling variants. The categories were identified by drawing on a range of existing taxonomies which were themselves compiled with reference to various written and electronic texts (e.g. J. Androutsopoulos 2000; D. Crystal 2003; R.-M. Weber 1986; T. Shortis 2007), and

adapted to account for the spellings seen in CorTxt. As will be described below, categorisation remains a somewhat subjective process, in which the views of the researcher are to some extent imposed on the data.

The functional categories of spelling variants which we identified using DICER are given in Table 5 along with examples and the frequency of each category in terms of tokens and types (percentages are out of the number of category assignments). Note that a substantial number of forms are placed in more than one category: <2nite>, for example, includes both a number homophone <2> and the eye dialect spelling of *night* <nite>.

At first glance, the functional categories may seem primarily motivated by the maxim of brevity and speed. For example, representing *you* with either <u> and <ya> involves abbreviation in the sense of fewer letters. However, the very fact that texters can choose between <u> or <ya> suggests that motivations other than

Table 5. Category examples and frequencies

Category	Examples	Tokens	Types
Letter homophones	u, r, ur, c, b	1040 (29.91%)	30 (3.25%)
Number homophones	person2die, 2gether, up4that,in2hospital, 2nite	476 (13.69%)	126 (13.64%)
Clippings	tomo, tho, v, bout, prob, hav	414 (11.91%)	113 (12.23%)
Apostrophe omission	wots, im, il, its, thats	367 (10.56%)	60 (6.49%)
Eye dialect	bak, luv, wots, gud	243 (6.99%)	47 (5.09%)
Colloquial contractions	lookin, av, cos, n, whaddya	232 (6.67%)	94 (10.17%)
Spacing	Thankyou, ur, u2, aswell, Ohdear, sleep4aweek	232 (6.67%)	171 (18.51%)
Consonant writing	txt, msg, lv, wld, pls	130 (3.74%)	51 (5.52%)
Mistyping	your (for you're), definately, adn, menas	61 (1.75%)	47 (5.09%)
Double letter reduction	stil, worry, spel, I'l, 2moro, ul	43 (1.24%)	16 (1.73%)
Misspelling	your (for you're), definately,	32 (0.92%)	26 (2.81%)
Unclear	ur = your, tomoz = tomorrow	31 (0.89%)	21 (2.27%)
Other abbreviations	no, happng, checkd, 2morw	29 (0.83%)	17 (1.84%)
Possible regional respellings	summat, summort, sumfing, dis	28 (0.81%)	8 (0.87%)
Predictive texting 'mistake'	in (for go), he (for if)	6 (0.17%)	2 (0.22%)
Visual morphemes	I'm@my; Lunch@12	5 (0.14%)	5 (0.14%)
No category assigned		108 (3.11%)	90 (9.74%)
Total		3477	924

brevity and speed may also be involved. In this case, the choice of <ya> (rather than <you> or <u>) represents an informal, spoken version of *you*, while the potential effects of <u> rest on the visual appearance of the word. The fact that spelling variants can be categorised into functional categories suggests meaningful and deliberate usage, a point supported by the relative infrequency of forms categorised as misspellings or mistypings.

By far the most frequently occurring category, in terms of tokens, is that of letter homophones, which account for nearly a third of all spelling variants categorised by DICER, and which include <u> (*you*), <r> (*are*), (*be*) and <c> (*see*). It is important to note, however, that the first homophone, <u>, accounts for the overwhelming majority of letter homophones. This may reflect the frequency of the second person pronoun; *you* is the most frequent word in the corpus, with 7,884 occurrences, 3,043 of which are spelt <u> (C. Tagg 2009: 135). In other words, the frequency of this category of spelling variation may reflect the occurrence of one particular lexical item as much as it does a general tendency towards using letter homophones, and the lower frequency in terms of types provides more evidence for this view. Similarly, number homophones are the second most frequent category (albeit with half the occurrences of letter homophones), and these chiefly comprise the numbers <2> and <4>. Again, this may be because of the frequency of these lexical items (*to* and *for*) which occur respectively in 2nd and 9th position in the word frequency list (and of course <2> can also represent *too*) (C. Tagg 2009: 219). What is also interesting about the number homophones is the number of times that they occur in alphanumeric sequences such as *up-4that*, that is, where spaces are omitted around the homophone. The omission of spacing is a frequent occurrence throughout the corpus. Omitting spacing around number homophones rather than, say, letter homophones, is probably a strategy motivated by intelligibility. The high frequency of these alphanumeric sequences also explains why the frequency of this category in terms of types is at a similar percentage to that of tokens, unlike for letter homophones. The high frequency of <2> and <4> certainly increases the category token frequency, but the variety of words which surround <2> and <4> accounts for the high type frequency. In the light of the media portrayal, it is interesting to note that there are only 4 occurrences of <8> as a homophone in the sample, all in the sequence *l8r*. As mentioned at the start of this paper, <8> appeared from our selected extracts to be a feature popularly associated with SMS (as in Humphrys' 'I h8 txt msgs').

The next most frequent category is clippings, accounting for nearly 12% of the respellings. A clipping is defined by D. Crystal (2003: 120) as 'part of the word which serves for the whole, such as *ad* and *phone*': that is, where either the beginning or end of the word has been clipped (or the middle in cases such as <spectacles> clipped to <specs>). One could speculate that these, at least where the end

of the word is clipped, may have become more common with the advent of predictive texting. The dictionary would not automatically recognise <tmwr>, but by four presses it would have predicted that the texter intended to write *tomorrow* and would suggest <tom>. Consonant writing is in fact much less frequent in the corpus. Whether speed would be achieved alongside brevity depends on whether predictive texting is used, and whether the vowel-less form had been entered in the dictionary. If predictive texting was used, and the form was not in the dictionary, it is likely that the variant form would involve more key presses than the 'standard' form. The assumption here, of course, is that brevity and speed are the motivations.

With colloquial contractions and eye dialect, both of which account for just under 7% of all spelling variant tokens, texters appear to be combining the maxim of brevity and speed with that of phonological representation; that is, they are capturing informal spoken forms. To take the (slightly) less frequent of the two first, colloquial contractions were defined by R.-M. Weber (1986:415) as 'spelling patterns that are regularly used to represent reduced, colloquial speech'. These include <kinda>, <ya>, <lookin> and <cos>. By definition, these forms involve abbreviation. However, the spellings also capture informal spoken forms. The other category, eye dialect, is in fact more complex than a simple representation of spoken forms. Respelt forms such as <gud> and <wots> do not reflect how words would be pronounced in fast, informal speech but represent a more common sound-symbol relationship (for example, <o> for the vowel sound in *what*). The effect of these is often to suggest an uneducated, ignorant speaker (D. Preston 2000). In SMSes, then, although the forms are generally shorter, the motivation for their use may lie in conveying a certain kind of identity (rebellious, careless, unconcerned) rather than simply an attempt to reduce characters. If colloquial contractions and eye dialect can be described as indexical, the implication is that homophones, clippings and consonant writing may similarly index identities, albeit in a way less directly associated with speech.

Before moving any further down the list of categories, it is useful to pause and consider how cautiously the above assertions should be made. As mentioned earlier, identifying functional categories and assigning certain forms to them relies to an extent on our assuming what texters were doing or what they thought they were doing. In some cases, forms could have been placed in more than one category, and the category that we chose to assign it to is therefore arguable. These include the examples shown in Table 6 (in each case, the chosen category comes first).

Some of these raise wider issues. To what extent can regional representations be distinguished from 'standard' colloquial contractions? The spellings <sum-mort> and <sumfing> seem fairly marked forms particular to certain dialects (e.g. West Country and London, respectively); as does <dis> for *this*. The rendering of *with* as <wiv> however, is more contentious, as is the reduction of <ing> to <in>

Table 6. Examples where category choice was particularly debatable

Spelling	Category Choice	Normalisation
	letter homophone — or clipping?	(be)
<n>	colloquial contraction — or clipping?	(and)
<tho>	clipping — or eye dialect?	(though)
<prob>	clipping — or colloquial contraction?	(probably and problem)
<bout>	clipping — or colloquial contraction?	(about)
<bak>	eye dialect — or abbreviation?	(back)
<lookin>	colloquial contraction — or clipping?	(looking)
<wiv>	colloquial contraction — or regional spelling?	(with)
<checkd>	other abbreviations — or consonant writing?	(checked)
<happng>	other abbreviations — or consonant writing?	(happening)
<tomoro>	clipping — or eye dialect?	(tomorrow)

and the h-dropping in <av> (for *have*). There is likely no absolute dividing line between these two categories. It could also be argued that ‘double letter reductions’ such as <stil> and <spel> are phonetic representations and could be categorised as eye dialect.

Some spellings were more difficult to put into any category at all. These included <tomoz> (for *tomorrow*) and <ur> (for *your*). The latter is particularly interesting. The spelling <ur> is used throughout the corpus to mean either *you are* or *your*. The former can be understood as a pair of letter homophones (<u> for *you* and <r> for *are*). This explanation does not hold for *your* represented as <ur>. A more accurate explanation may be that <ur> is a clipping of *your*. However, this explanation feels less satisfactory than a phonetic one, perhaps given the phonetic use of <ur> elsewhere. If <ur> represents *you are*, and therefore *you’re*; then it seems a small step to representing *your* as <ur>. Given the uncertainty, this variant was placed in the Unclear category (as was <tomoz>).

The results reveal very few forms which were deemed to be ‘mistakes’, either mistypings (such as <adn> for *and*) or misspellings (such as <definately>). This has some implications for the category of apostrophe omission, which occurs very frequently across the corpus. Given that spelling variation elsewhere in the corpus is seemingly so deliberate and principled, one could argue that apostrophe omission may similarly constitute deliberate attempts to either save time, space or effort, or to create a certain effect. In the case of <wots> for example, the lack of apostrophe adds to the eye dialect to create a sense of rushed, casual speech. The main point to emerge from the relative lack of misspellings and mistypings is that spelling practices can be described as deliberate and meaningful.

There are of course a number of problems with the above assertion. One question that emerges is the extent to which ‘mistakes’ and deliberate respellings can be distinguished. Our starting assumption was that spelling variation is functional, principled and meaningful and this inevitably shaped our categorisation. The spelt forms <wot>, <wory> and <hav> were therefore labelled as the deliberate spelling strategies of, respectively, eye dialect, double letter reduction and clipping. Whether or not any instances of these forms were genuine mistakes is hard to tell. Coming at this from the opposite direction, those that were classified as ‘mistakes’ were deemed so because they did not fit the categories we had identified — they included transpositions such as <adn> (*and*) and <menas> (*means*), as well as <your> (*you’re*) and <definately> (*definitely*). However, given a different categorisation system, who is to say that transpositions could not take on some distinct meaning? The classification process is somewhat circular. This is also well illustrated by the omission of apostrophes, which occurred on numerous occasions throughout the corpus. To an extent, <wots>, <im> and <thats> can be seen as part of the intimate, casual and playful language of Txt, and one could argue that it is unlikely that the people who contributed to the corpus would have mistakenly omitted the apostrophe in, say, *that’s*; but where is the line drawn? Confusion between *it’s* and *its* is arguably fairly widespread and such confusion is likely to occur in this corpus.

5.2 Summary of the DICER categories analysis

Bearing in mind the caveats highlighted, certain conclusions can be drawn from the DICER analysis. Our study confirms and extends R. Grinter and M. Eldridge’s (2003) observation that spelling variants tend to be of common words — they suggest *tomorrow* and *homework* among their teenage texters. Our findings reveal that very frequent grammatical words are in fact those that have their spelling most commonly altered. The most frequently varied words are *you* and *to*, and this explains the predominance of letter and number homophones <u> and <2>. In other words, a few, very commonly used and regularly respelt words account for a great deal of the variation in spelling in the corpus. On the other hand, missing are the long abbreviated phrases listed by D. Crystal (2008): *AFAIK*, *ASLMH* and so on. What this suggests is a different set of practices than that described in the popular press; practices which emerge from and facilitate interaction.

The identification of functional categories facilitated by DICER show that patterns of spelling variation can be categorised across large datasets. The implication is that spelling variation is motivated not only by the need for brevity and speed. As C. Thurlow and A. Brown (2003) note, spelling variation is also motivated by the communicative demands of the medium as well as the attempt to index

informal and casual social identities through the representation of spoken forms (such as colloquial contractions and eye dialect). However, the indexical potential of respelling does not appear to lie only in its relation to spoken forms. Letter and number homophones, clippings and consonant writing play instead with the visual form of the word. The distribution of frequencies across the categories in our study suggests the visual may play a much larger role than attempts to reflect speech.

Spelling variants that do not fit functional categories are surprisingly few (given concerns in the print press regarding the chaotic and damaging nature of texting), thus offering some support to the general suggestion that spelling variants are deliberate and meaningful. Overall, what emerges is that SMS spelling variants *can* be categorised across a large dataset not only according to formal patterns but also to functional categories including colloquial contractions and eye dialect.

Conclusions

In this paper, we have reported on the use of VARD2 and DICER to manually normalise and analyse spelling variation in a corpus of SMS text messages, CorTxt. Our aim was to contribute to a description of the spelling variation that occurs in text messaging with findings based on a significant quantity of empirical data. We have highlighted various spelling trends found in the corpus which expose the difficulty of automatic normalisation and uncovered specific spelling rules which allowed us to categorise the spelling variants. Our analysis suggests that a few frequently occurring forms, chiefly the homophone <u>, account for a great deal of the spelling variation, and that these are used alongside colloquial contractions and eye dialect in playful identity construction. While our research provides valuable empirical evidence for the ideas about texting being advanced by linguists, it should be noted that the findings apply to the spelling practices of an educated, literate network and that younger or less educated texters may diverge more widely from the orthographic principles described above. Further research is needed to explore the extent to which this is the case.

As detailed in the article, we were able to build up a set of spelling rules which could subsequently be used to automatically normalise spelling across larger SMS corpora. The caveat here, of course, is that spelling variation practices may differ slightly across different communities and thus datasets. In an initial study (not detailed here) we find that these rules are useful when used within VARD and allow for a substantial amount of automatic normalisation of CorTxt after training. Future work could build upon this initial study to assess how useful a spelling rule based method could be for normalising SMS corpora.

The method of spelling variation analysis described here, *i.e.* using VARD and DICER, could also be applied to other language varieties which contain substantial spelling variation — both for finding spelling trends and categories and for finding spelling rules which can be used for automatic normalisation. For example, the progress made in automatically normalising Early Modern English spelling variation (see A. Lehto *et al.* 2010) can be improved further with an improved rule base (A. Baron *et al.* 2009a) and other forms of computer-mediated communication (CMC, e.g. Twitter) corpora could have their spelling variation analysed, with trends and categories found and comparisons made between CMC varieties.

Acknowledgments

Alistair Baron's and Paul Rayson's contribution to this paper was co-funded by the EPSRC and ESRC (EP/F035438/1, EP/F035071/1, EP/F035454/1), Project title "Isis: Protecting children in online social networks".

References

- Androutsopoulos, Jannis K. 2000. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics* 10(4): 419–438.
- Anis, Jacques. 2007. Neography: Unconventional Spelling in French SMS. In *The Multilingual Internet: Language, Culture, and Communication Online*, B. Danet, S.C. Herring (eds.), 87–115, Oxford: Oxford University Press.
- Archer, Dawn; Tony McEnery; Paul Rayson; Andrew Hardie. 2003. Developing an automated semantic analysis system for Early Modern English. In *Proceedings of Corpus Linguistics 2003*, D. Archer, P. Rayson, A. Wilson, T. McEnery (eds.), Lancaster, UK: Lancaster University.
- Baron, Alistair; Paul Rayson. 2009. Automatic standardisation of texts containing spelling variation: How much training data do you need? In *Proceedings of Corpus Linguistics 2009*, M. Mahlberg, V. González-Díaz, C. Smith (eds.), Liverpool: University of Liverpool.
- Baron, Alistair; Paul Rayson; Dawn Archer. 2009a. Automatic standardization of spelling for historical text mining. In *Proceedings of Digital Humanities 2009*. Maryland, USA: University of Maryland.
- Baron, Alistair; Paul Rayson; Dawn Archer. 2009b. Word frequency and key word statistics in historical corpus linguistics. *International Journal of English Studies* 29(1): 41–68.
- Beaufort, Richard; Sophie Roekhaut; Louise-Amélie Cougnon; Cédric Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Sweden: Uppsala.
- Church, Kenneth W.; William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing* 1(2): 93–103.
- Cook, Vivian. 2004. *Accommodating Broccoli in the Cemetery: Or Why Can't Anybody Spell?* New York: Touchstone.

- Crystal, David. 2003. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.
- Crystal, David. 2008. *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.
- Deumert, Ana; Sibabalwe O. Masinyana. 2008. Mobile language choices — The use of English and isiXhosa in text messages (SMS): Evidence from a bilingual South African sample. *English* 29(2): 117–147.
- Döring, Nicola. 2002. “1 bread, sausage, 5 bags of apples I.L.Y” — communicative functions of text messages (SMS). *Zeitschrift für Medienpsychologie* 3.
- Fairon, Cedric; Sébastien Paumier. 2006. A translated corpus of 30,000 French SMS. In *Proceedings of LREC 2006*, Genova.
- Görlach, Manfred. 1991. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.
- Grinter, Rebecca E.; Margery Eldridge. 2003. Wan2tlk?: Everyday text messaging. In *Proceedings of the CHI'03 Conference on Human Factors in Computing Systems*. New York: ACM.
- Hård Af Segerstad, Yvla. 2002. *Use and adaptation of written language to the conditions of computer-mediated communication*. PhD thesis, Department of Linguistics, Göteborg University, Sweden.
- Hardy, Thomas. 1891. *Tess of the d'Urbervilles*. London: Random House.
- Kasesniemi, Eija-Liisa; Pirjo Rautiainen. 2002. Mobile Culture of Children and Teenagers in Finland. In *Perpetual Contact: Mobile Communication, Private Talk, Public Performance*, J. Katz, M. Aakhus (eds.), Cambridge: Cambridge University Press.
- Kessler, Angela; Alexander Bergs. 2003. Literacy and The New Media: Vita Brevis, Lingua Brevis. In *New Media Language*, J. Aitchison, D. Lewis (eds.), London: Routledge.
- Lehto, Anu; Alistair Baron; Maura Ratia; Paul Rayson. 2010. Improving the Precision of Corpus Methods: The Standardized version of Early Modern English Medical Texts. In *Early Modern English Medical Texts: Corpus Description and Studies*, I. Taavitsainen, P. Pahta (eds.), Amsterdam: John Benjamins.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting insertions, deletions and reversals. *Cybernetics and Control Theory* 10(8): 707–710.
- Ling, Rich; Naomi S. Baron. 2008. Text messaging and IM: Linguistic comparison of American college data. *Journal of Language and Social Psychology* 26(3): 291–298.
- Ling, Rich. 2005. Mobile Communications Vis-à-Vis Teen Emancipation, Peer Group Integration and Deviance. In *The Inside Text: Social Perspectives on SMS in the Mobile Age*, R. Harper, A. Taylor, L. Palen (eds.), Vol. 4, *The Kluwer International Series on Computer Supported Cooperative Work*. Amsterdam: Springer.
- Mitton, Roger. 1987. Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management* 23(5): 495–505.
- Mitton, Roger. 2008. Ordering the suggestions of a spellchecker without using context. *Natural Language Engineering* 15(2): 173–192.
- Palander-Collin, Minna; Mikko Hakala. 2011. Standardizing the Corpus of Early English Correspondence (CEEC). Poster presented at ICAME 32, Oslo, 1–5 June 2011.
- Pedler, Jennifer; Roger Mitton. 2010. A large list of confusion sets for spellchecking assessed against a corpus of real-word errors. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis, M. Rosner, D. Tapias (eds.), European Language Resources Association (ELRA), Valletta, Malta.
- Pound, Louise. 1925. The Kraze for “K”. *American Speech* 1(1): 43–44.

- Preston, Dennis R. 2000. Mowr and mowr bayud spellin': Confessions of a sociolinguist. *Journal of Sociolinguistics* 4(4): 614–621.
- Rayson, Paul; Alistair Baron. 2011. Automatic Error Tagging of Spelling Mistakes in Learner Corpora. In *A Taste for Corpora: In Honour of Sylviane Granger*, F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.), Amsterdam: John Benjamins.
- Sebba, Mark. 2007. *Spelling and Society: The Culture and Politics of Orthography around the World*. Cambridge: Cambridge University Press.
- Shortis, Tim. 2007. Revoicing Txt: Spelling, Vernacular Orthography and 'Unregimented Writing'. In *The Texture of Internet: Netlinguistics in Progress*, S. Posteguillo, M.J. Esteve, M. L. Gea-Valor (eds.), Newcastle: Cambridge Scholars Publishing.
- Tagg, Caroline. 2009. A corpus linguistic study of SMS text messaging. Unpublished PhD thesis, English Department, University of Birmingham (March 2009).
- Tagg, Caroline. 2012. *The Discourse of Text Messaging: Analysis of SMS Communication*. London: Continuum.
- Thurlow, Crispin; Alex Brown. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* 1(1).
- Thurlow, Crispin. 2006. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication* 11: 667–701.
- Weber, Rose-Marie. 1986. Variation in spelling and the special case of colloquial contractions. *Visible Language* 20(4): 413–426.
- Yvon, Francois. 2010. Rewriting the orthography of SMS messages. *Natural Language Engineering* 16(2): 133–159.

Summary

Spelling variation, although present in all varieties of English, is particularly prevalent in SMS text messaging. Researchers argue that spelling variants in SMSes are principled and meaningful, reflecting patterns of variation across historical and contemporary texts, and contributing to the performance of social identities. However, little attempt has yet been made to empirically validate SMS spelling patterns (for most languages, with the notable exception of French) and verify the extent to which they mirror those in other texts.

This article reports on the use of the VARD2 tool to analyse and normalise the spelling variation in a corpus of over 11,000 SMSes collected in the UK between 2004 and 2007. A second tool, DICER, was used to examine the variant and equivalent mappings from the normalised corpus. The database of rules and frequencies enables comparison with other text types and the automatic normalisation of spelling in larger SMS corpora.

As well as examining various spelling trends with the DICER analysis it was also possible to place the spelling variants found in the SMS corpus into functional categories; the ultimate aim being to create a taxonomy of SMS spelling. The article reports on the findings from this categorisation process, whilst also discussing the difficulty in choosing categories for some spelling variants.

Lol, mdr and ptdr

An inclusive and gradual approach to discourse markers

Deniz Uygur-Distexhe

Université catholique de Louvain, Institut Langage et Communication,
VALIBEL

Introduction

Text messaging is a spontaneous computer-mediated communication type which is asynchronous (there is a delay between the reception and the sending of an SMS) and monomodal (“it employs writing as means of expression” Y. Hård af Segerstad 2005: 34). Texting shows graphic immediacy (*immédiat graphique* P. Koch and W. Oesterreicher 2001: 614) parameters and has writing constraints linked to the tool (space, cost, time, encoding) and the communicative situation (physical channel, cultural use, communication purpose, relation between participants, relations to external context and to text, following D. Biber 1988: 39).

Text messaging has given rise to a “*new language*¹ of abbreviations, acronyms, word combinations and punctuations” (C.K. Varnhagen *et al.* 2010: 719). C.C. Werry (1996) and C.K. Varnhagen *et al.* (2010) hypothesize that the main purpose of the participants who created these short cuts to express “words, phrases, and emotions as well as textual and graphical pragmatic devices” (2010: 720) was to increase the speed of the communicative exchange. C.K. Varnhagen *et al.* (2010: 720) add that because of their ubiquity, some of these short cuts and pragmatic devices now have entries in dictionaries (E. Jansen, 2003; <http://www.netlingo.com>; <http://www.urbandictionary.com>). Typical of computer-mediated communication as extended to casual conversation, *LOL* (*Laughing Out Loud*) appeared in the *Oxford English Dictionary* in March 2011 (<http://www.oed.com/>

1. One of the reviewers notices that the term language is debatable and refers to R. Panckhurst (2009) for usage of “écriture SMS” *i.e.*, text ‘writing’ as opposed to ‘language’ and to L.-A. Cougnon and G. Ledegen (2008) for usage of “écrit SMS”.

viewdictionaryentry/Entry/291168). Borrowed from English, French *lol*² appears in the 2013 edition of *Le Petit Robert* (see below).

LOL interj. – 2000. *acronyme anglais, de laughing out loud, « éclaté de rire, mort de rire », d’abord dans les SMS. Interjection soulignant le caractère comique, l’ironie d’un propos (surtout écrit). Tu vas te marier, toi ? lol !*

(*LOL interj.* – 2000. English acronym, from *laughing out loud*, “to split one’s sides with laughter, to be helpless with laughter”, initially in SMSes. Interjection underlining the comic aspect or irony of a comment (mostly written). *You’re getting married? lol!*)

Other initialisms such as *MDR* (*Mort De Rire* — literally ‘died from laughing’ or ‘helpless with laughter’) and *PTDR* (*PéTé De Rire* — literally ‘burst from laughing’ or ‘to split one’s sides with laughter’) do not yet have entries in general dictionaries, although, according to N. Michot (2007: 3), they are used on a massive scale in computer-mediated communication, but of course it may also be due to the fact that these initialisms are abbreviations rather than acronyms.

Given the extensive use of *lol*, *mdr* and *ptdr* compared to other initialisms, this paper aims to study how these three initialisms work. Considered mainly as interjections, I hypothesize that *lol*, *mdr*, and *ptdr* could work syntactically (through their non-integration in the syntactic clause and position) and semantically (through their semantic opacity and procedural meaning) as discourse markers. Because of the SMS corpus on which it draws and its focus on *lol*, *mdr* and *ptdr*, this study is distinct from the huge quantity of research on ‘more classical’ discourse markers, also known as pragmatic markers or discourse particles, such as *well*, *you know*, *I mean*, through oral and written corpora in the last thirty years (for a general survey of research on discourse markers, see K. Aijmer and A.-M. Simon-Vandenberg 2009). In addition, this paper excludes the typical sociolinguistic perspective adopted in works on *lol* (see S.A. Tagliamonte and D. Denis 2008), taking instead an approach based on discourse analysis.

From this perspective, I focus on the relationship between the different syntactic and semantic characteristics of *lol*, *mdr* and *ptdr* as discourse markers. It is assumed that *lol*, *mdr* and *ptdr* can be considered as discourse markers if the term *discourse marker* is interpreted in a gradual and inclusive sense. For the inclusive aspect, the term *discourse marker* was deliberately chosen among several others, such as *discourse connector*, *discourse operator*, *pragmatic connector* and *sentence connector*, since *discourse marker* is the most often used designation

2. *Lol*, *mdr* and *ptdr* can be written in upper case or lower case, as explained for *LOL/Lol* in the Oxford English Dictionary. In this paper, the use of upper case is limited to initialisms accompanied by a reference to the words composing them.

(K. Aijmer and A.-M. Simon-Vandenberg 2009; K. Fischer 2006; L. Schourup 1999), the most inclusive category (K. Aijmer and A.-M. Simon-Vandenberg 2009; K. Fischer 2006) and the most functional name (K. Fischer 2006). The discourse marker works as a signpost, providing instructions on how to interpret the message (K. Aijmer and A.-M. Simon-Vandenberg 2009). This approach to discourse markers is also inclusive because it considers discourse markers as covering pragmatic markers as defined by B. Fraser (2006: 189).

Through the analysis of *lol*, *mdr* and *ptdr* as discourse markers, this work aims to provide empirical access to SMS as a genre (D. Biber 1988), by showing that involvement (W. Chafe and J. Danielewicz 1987: 108) is typical of SMS language since messages are exchanged between people who already know each other and that the sender trusts the receiver's capacity to make pragmatic inferences in order to decode the message (Y. Hård af Segerstad 2005: 34). In this paper, I analyse a French text message corpus collected in France and Belgium. The choice to analyse a text message corpus is driven by its informal language style, which is known to favour the production of discourse markers (L. Schourup 1999: 234).

Before proceeding to the analysis of *lol*, *mdr* and *ptdr*, I would first like to consider the literature on these three initialisms and to distinguish their uses. Secondly, I will describe the syntactic and semantic characteristics of discourse markers as well as the value of taking into account their context and characterizing their speech functions. Thirdly, I will outline the characteristics of text messaging, a type of computer-mediated communication. Fourthly, I will present the data and the methodology I use for this work.

1. Literature on *lol*, *mdr* and *ptdr*

Lol, *mdr* and *ptdr* are synonyms, but, interestingly, most French users do not know the exact meaning of the English initialism. So French speakers/texters often use an approximate translation of *lol*, such as 'to laugh' or 'funny' (N. Michot 2007: 7). In addition, some senders use another referent for *LOL*, that is *Lots Of Love*. This could create awkward situations, as in the famous SMS sent by a mother who wrote to her son *your great aunt just passed away. LOL*³

Mostly pronounced as an acronym, *lol* can be used as a lexeme and preserves a single form whatever its syntactic category (N. Michot 2007: 7). In the selected SMS corpus for this study, *lol* can work as a verb (1) or as an adjective (2).

3. This example is available on <http://www.uselesshumor.com/2011/03/funny-texts-your-great-aunt-just-passed.html>

- (1) *chanj pa(kwak tet1pe...Jlol!)*⁴
(don't change (well, maybe a little...I lol!))
- (2) *Slt je shte une super LOL soirée a tt l'équipe de ce test!*
(Hi I wish a super *lol* party to the whole team of this test!)

Mdr and *ptdr* are not pronounced as acronyms. Like *lol*, they are characteristic forms of laughter. They all mean 'ha ha!' and are used to draw attention to a joke or humorous statement, or to express amusement. Contrary to the acronym *lol*, the status of *mdr* and *ptdr* seems fixed as initialisms, although their phrasal status of adjectival phrase can be underlined by the use of a copula, as in (3).

- (3) *T'es une vraie garce j'adore !! Chuis mdr :)*
(You are a real bitch I love it! I'm *mdr*)

Lol, *mdr* and *ptdr* could be classified as emotional language, but they are not restricted to a filler function as S.A. Tagliamonte and D. Denis (2008) claim. They consider *lol* as a "signal of interlocutor involvement" used by "participants in the flow of conversation" (S.A. Tagliamonte and D. Denis 2008: 11), also known as a "backchannel signal". Among other functions, *lol* can be a "phatic filler comparable to *okay*, *really*, or *yeah* in spoken discourse" (N. Baron 2004, 416 quoted in S.A. Tagliamonte and D. Denis 2008: 29): it can express a phatic function in language in the sense of R. Jakobson (1960), as C.K. Varnhagen (2010: 729) claims. In addition, following N. Michot (2007: 7), *lol* can be used to punctuate or to qualify in an ironic way when *lol* undergoes a semantic inversion effect and produces a proximity effect thanks to its positive polarity. N. Michot (2007: 7-8) adds that *lol* suffers from semantic bleaching and is used in the same manner as smileys, which represent the state of mind of the writer. In this paper, it will be shown that *lol*, *mdr* and *ptdr* can also be considered as discourse markers.

Lol, *mdr* and *ptdr* are originally and chiefly used in the language of electronic communication, in a colloquial and slang-like way.⁵ While the expansion of the use of *lol* in French is quite recent, the question is whether *lol*, *mdr* and *ptdr* occur simply in contemporary slang or whether they are involved in the general process

4. I have chosen to use the original form of the SMS found in the SMS corpus selected for this work, a corpus that I will describe in the section on data and methodology. In order to help readers who do not know French, the English translations are not literal.

5. Le Robert 2013 provides the first evidence of *lol* in French in text messaging in 2000, while the Oxford English Dictionary dates the first occurrence of English *lol* back to 1989, in an electronic journalistic text. Another difference between the entry for *lol* in the Oxford English Dictionary and in Le Robert 2013 is that the former defines *lol* as an interjection or a noun while the latter restricts the uses of *lol* to interjection only.

of language evolution, as some researchers such as C. Lewis and B. Fabos (2005), G. Merchant (2001), J.N. Spatafora (2008) or S. A. Tagliamonte and D. Denis (2006) seem to hypothesize. To sum up, *lol*, *mdr* and *ptdr* are ubiquitous in computer-mediated communication texts, but *lol* is the only initialism to have been included in general dictionaries such as the Oxford English Dictionary (OED) or Le Petit Robert (Robert).

In order to show how I analyse these three initialisms, I will first describe the main syntactic and semantic characteristics accepted as defining discourse markers. Next, I will show the specificity of text messaging and its influence in the use of discourse markers. Lastly, I will present the corpus and the methodology used to analyse *lol*, *mdr* and *ptdr* as discourse markers.

2. Syntactic and semantic characteristics of discourse markers

According to D. Vincent (2005: 189), the expression *discourse marker* can be applied to “any element of language that plays a role in the organization of discourse”. Discourse markers are “distributed around two functional poles: connectors — which ensure the articulation of utterances and, therefore, function like conjunctions and adverbs of liaison — and modal elements — which introduce a point of view on the discourse and take the form of adverbial expressions” (D. Vincent 2005: 189). This functional distribution made me wonder whether other functional poles exist that may characterize specific subclasses of discourse markers.

As I explained in the introduction to this paper, I have chosen a gradual and inclusive approach to the definition of discourse markers. My approach is gradual in that I focus on the syntactic and semantic characteristics of discourse markers. This enables a large number of linguistic elements to be taken into account as discourse markers.

Broadly speaking, discourse markers could be any linguistic element which is multifunctional, which relates to the communicative situation, and which expresses attitudes and emotions (C. Bazzanella 2006: 449). I also loosely follow the definition suggested by D. Lewis (2006: 44), who states that discourse markers are expressions that combine (i) “the semantics of discourse-relational predications” (such as the attitudes of the speaker) with (ii) “syntactic dependency on a clausal host” and expressions with (iii) “low information-salience”.

From a syntactic and semantic perspective, my approach is inclusive because my definition of discourse markers is not couched in negative terms, as is the case with many definitions of discourse markers, criticized by D. Lewis (2006: 44). These negative characterizations are often used to define discourse markers as non-propositional, with non-truth-conditional meaning, having little or no semantic

value, placed outside the syntax of the sentence, optional, etc. My approach is also gradual because the different syntactic and semantic criteria are not binary and always leave room for doubt. In my syntactic approach, discourse markers are not integrated into the syntactic clause, but this does not mean that they are outside the syntax. Discourse markers may be isolated from the sentence, but they also occur in other positions in the sentence such as the initial, the quasi-initial, the medial and the final positions. My semantic approach is characterized by two gradual criteria: semantic opacity and procedural meaning. Discourse markers can be used in figurative or opaque meanings and can work in a procedural rather than a conceptual way (in a conceptual relation linguistic elements specify “a defining set of semantic features”, whereas in a procedural relation, they specify how the segments they introduce are “to be interpreted relative to the prior, subject to the constraints mentioned earlier” according to B. Fraser 1999: 944). Semantically, I leave a place for lack of knowledge concerning the referent of the discourse marker or the clauses that the discourse marker links. This flexible way of applying these criteria explains how my approach is gradual. I do not impose the obligation that all the criteria should be met simultaneously in order for a specific example to be counted as a discourse marker. Above all, I pay attention to the degree to which a discourse marker fulfils syntactic and semantic criteria.

Since *lol*, *mdr* and *ptdr* are first considered as interjections, my aim is to show that these interjections form part of the discourse marker category. If interjections are viewed as a sub-category of discourse markers, according to Schiffrin (1987: 73-101), they have to be hosted in a sentence and not isolated (otherwise I have to prove that isolated interjections function in the same way as discourse markers).

With this inclusive and gradual approach, by assigning scores for the syntactic and semantic parameters defining discourse markers, I rank occurrences of *lol*, *mdr* and *ptdr* according to the sum of their discourse marker scores. Instances of *lol*, *mdr* and *ptdr* whose score qualifies them as discourse markers are then analysed functionally, following the moves described by S. Eggins and D. Slade (1997: 192). In order to analyse the moves, I take into account the mood of the host sentence. In addition, I consider the co-occurrences with other discourse markers. The functional analysis aims to understand the variety of uses the discourse markers *lol*, *mdr* and *ptdr* can have.

Before discussing methodology, I will present the specificity of text messaging and its influence on the use of discourse markers.

3. Text messaging

Text messaging is first and foremost a written medium, just like all text-only CMC. This means that users “do not see or hear their interlocutors and thus do not have access to non-verbal information about how others are responding” (S.C. Herring 1999). On the other hand, texting also has oral characteristics in the way the writer is involved with the recipient (see W. Chafe 1982 for the notion of “involvement”) and uses a list of colloquial expressions such as *lol*, *mdr* and *ptdr* that show this involvement. According to D. Biber (1988: 24), it is difficult to define the relations between spoken and written texts since these relations are complex “and associated with a variety of different situational, functional, and processing considerations”.

This monomodal medium is known to be a spontaneous computer communication type with casual spoken texts, but short formal messages can also be written. Among the criteria used to define computer-mediated communication, for example chatting and e-mailing, there are some that are used to define text messaging: the conception delay (the writing of a message takes time and is not immediate), asynchronicity, *i.e.* “there is often a considerable time lag between when a message is sent and when it is responded to” (S.C. Herring 1999), and unilaterality, *i.e.* there is no turn interruption if we accept that one SMS is equal to one turn.⁶

In the writing of SMSes, there are different strategies to produce expressivity, proximity or familiar exchange, and graphic space economy. Among these strategies, pragmatic devices such as discourse markers or smileys can be used. For example, smileys could work as a “substitute for non-referential communication in face-to-face exchanges” (C.K. Varnhagen *et al.* 2010: 729). These pragmatic devices could help users to clarify meaning. According to R. Jakobson (1960), emotion and attitude are expressed in face-to-face conversation by the means of “facial expressions, such as smiling and frowning”, but these means do not exist in text to make emotions explicit. In text, smileys and other pragmatic devices such as *lol* are adaptations to express emotions “without providing lengthy written explanations” (C.K. Varnhagen *et al.* 2010: 730).

To understand how *lol*, *mdr* and *ptdr* could function as discourse markers in texting, I manually analyse all the occurrences in my SMS corpus, which will be presented in the next section.

6. Each message is considered as a turn since *turn*, in synchronous computer-mediated communication, is “the term for an individual message of any length sent by a user” (S.A. Tagliamonte and D. Denis 2008: 28). In addition, an SMS is mostly sent to a single receiver, but in some cases there could be a group of receivers.

4. Data and methods

Before presenting the parameters chosen to define *lol*, *mdr* and *ptdr* as discourse markers, I will describe the creation and composition of the SMS corpus I use.

4.1 The SMS corpus

The SMS corpus I use is composed of two samples of 4,000 SMSes each, randomly selected⁷ from 4,491 SMSes gathered in France and from 65,935 SMSes gathered in Belgium. The purpose was to work on a balanced corpus of French SMSes from France and Belgium so as to avoid French regional variations as much as possible. The whole corpus is not available yet, but some of the French SMSes gathered in Belgium are accessible through the *SMS4science* project (<http://www.SMS4science.org/?q=en>)⁸ and a CD-ROM (C. Fairon *et al.* 2006).

The corpus of 8,000 text messages amounts to a total of 126,135 words and the average length of a message is about 20 words.⁹ The corpus is composed of single messages, *i.e.* messages are isolated from the discussion process.¹⁰ As a consequence, it is impossible to access the conversational exchange as a whole. This also explains why my analysis is confined to the micro-syntactic level.¹¹ In addition, there are a number of uncertain cases in which it is difficult to syntactically place the linguistic item or to analyse its meaning since the context is insufficient.

From this sample of 8,000 SMSes, I automatically selected all the messages containing *lol*, *mdr* and *ptdr*. Then I separated the messages into three dedicated Excel sheets: one per marker. Because some messages include more than one of these three initialisms, I duplicated them in order to count one occurrence for each message. I annotated 519 messages containing *lol*, *mdr* and *ptdr* according to the parameters chosen to define discourse markers, namely syntactic and semantic criteria. Then, in a prototypical approach,¹² I gave a score to each parameter

7. The random selection was performed in Excel with the 'rand formula'.

8. The SMSes gathered in France are also a part of *sms4science* project.

9. In my corpus, the length of an SMS could vary between one word and 43 words. However, there is a bias in my data: some of the messages have been cut because of Excel's limit of 255 characters (spaces included).

10. The details of the SMS collection are explained in the introduction of this volume.

11. Consequently, I do not take into account illocutionary units, as defined by C. Benzitoun *et al.* (2010).

12. This prototypical approach needs to be validated by comparing my analysis with "typical discourse makers" such as *tu vois* 'you see', *bon* 'well', *alors* 'then', which should receive high scores.

and summed these scores to obtain a ranking of the occurrences from the highest discourse marker score to the lowest (8 to 0). Finally, I selected the occurrences with a high score (4 or more) to annotate their context and their speech functions.

4.2 The discourse marker parameters

As mentioned in Section 2, I have tried to approach discourse markers in an inclusive and gradual way. This discourse marker definition has been restricted to syntactic and semantic criteria because these syntactic and semantic parameters are controversial enough to exclude pragmatic parameters that are quite difficult to set out.¹³

This approach is gradual and inclusive because the occurrences are annotated in degrees; there is a place for doubt and no obligation to satisfy the complete set of criteria in order to define *lol*, *mdr* and *ptdr* as discourse markers. Before presenting the ranking of discourse marker scores, I will first describe the syntactic parameters, then the semantic parameters.

4.2.1 Syntactic parameters

To distinguish *lol*, *mdr* and *ptdr* as discourse markers, I chose three syntactic parameters: integration in the syntactic clause, non-integration in the syntactic clause, *i.e.*, *lol*, *mdr* and *ptdr* positioned in the overall sentence, and grammatical category (irrespective of whether the occurrence is integrated in the syntactic clause).

As L. Brinton (1996: 34) highlights, discourse markers “occur either outside the syntactic structure or loosely attached to it and hence have no clear grammatical function”. I therefore looked at the integration of *lol*, *mdr* and *ptdr* in the syntactic clause. To detect this integration, I used the clefting process,¹⁴ such as *c’est lol que...* ‘it’s *lol* that’.

For initialisms that are not integrated in the syntactic clause and, more generally, for discourse markers, position is relevant. According to L. Brinton (1996: 34), although it is often said that discourse markers “are restricted to sentence-initial position”, it “is pointed out in general studies and shown in studies of individual markers” that “they frequently appear sentence medially and finally as well”. To examine this versatility, I determined five possible positions: the initial, quasi-initial, medial and final positions as well as an isolated location. Table 1 sets out the five positions with a list of parameters and examples to illustrate them.

13. This will be the object of further research.

14. For more explanation of the clefting process, see C. Blanche-Benveniste *et al.* (1984; 1991).

Table 1. Parameterized and illustrated positions

Position	Parameter	Example
Initial	First word position in a sentence.	(4) [...] <i>mai trouve l'endroi et l'hosi..</i>
	Can be preceded by a closing punctuation mark ¹⁵ such as a full stop, an exclamation mark, a question mark and even a smiley, but not followed by a closing punctuation mark, an upper case or nothing.	<i>Lol n'importe ou g m enfou mai trouve [...]</i> (but choose the place and also the time... <i>lol</i> anywhere I don't care but find)
Quasi-initial	Not the first word position but before syntactic clause. Can occur after an opening formula such as OK.	(5) <i>Ok lol tu me tien au jus</i> (Ok <i>lol</i> get back to me)
Medial	Located within the field governed by the verb.	(6) <i>Alor l'histoire lol enfet jsi po tro cki cp.c dc jsrai po juger ms entre ns el recom a dec grave</i> (So the story <i>lol</i> in fact I'm not quite sure what happened so I couldn't judge but between us she's started really messing around again)
Final	Last word position in a sentence. Not preceded by a closing punctuation mark such as a full stop, an exclamation mark, a question mark or even a smiley. Can be followed by a closing punctuation mark, a bracket, an upper case, closing formulas such as <i>à bientôt</i> 'see you soon', <i>à demain</i> 'see you tomorrow' or <i>bisous</i> 'kisses' or nothing.	(7) <i>Ah ben tu va voir lol</i> (Ok well you'll see <i>lol</i>)
Isolated	Not preceded or followed by a word. Can be preceded by a closing punctuation mark, bracket, smileys, and followed by a closing punctuation sign, bracket, smileys, upper case, or closing formulas.	(8) [...] <i>Et prépare maman car j'ai kk chose a lui demandé. Lol a tte [...]</i> (And prepare Mum because I have something to ask her. <i>Lol</i> see you soon)

While it is quite obvious that discourse markers “do not constitute a separate syntactic category” (B. Fraser 1999: 943), they also have different sources of grammatical categories. For the initialism *lol*, I categorized the occurrence as an

15. It is quite problematic to use punctuation to define syntactic position, particularly in texting where punctuation is unstable. To solve this issue, macro-syntactic analysis will be used in further research.

interjection in all cases of non-integration in the syntactic clause (when the initialism is separate from microsyntax) and as a verb or an adjective, as in the examples (1) and (2) presented above. Concerning *mdr* and *ptdr*, the literal French expressions *mort de rire* — literally ‘died from laughing’, or ‘helpless with laughter’ and *pété de rire* literally ‘burst from laughing’ or ‘to split one’s sides with laughter’ are clearly adjectival phrases, but when they are not integrated in the syntactic clause (because they cannot be clefted) they work as interjections,¹⁶ as in (9) for *mdr*.

- (9) *Au cas si vous faites autre chose dites moi toujours je verrai pour vous caser dans mon emploi du temps de princesse... Entre 2 bals mdr bon allez a +*
(If you’re doing something else tell me and I will see if I can fit you into my princess diary... between two balls *mdr* ok, see you soon)

These three syntactic criteria, namely integration in the syntactic clause, position in the sentence, and grammatical category, are parameters that also leave room for doubt. For each criterion, there are ambiguous cases:

- integration in syntactic clause or not for *lol* as in (10): if *lol* is integrated in the syntactic clause it means that ‘in that case, it is funny to be doomed to get bored at home the whole afternoon; if *lol* is not integrated in the syntactic clause, *lol* is the consequence of ‘being doomed to get bored at home the whole afternoon’.

(10) *J’sui condamnée a m’enuié ché moi toute l’après-midi alor lol*
(I’m doomed to get bored at home the whole afternoon then *lol*)

- initial or final positions for *lol* as in (11), *mdr* as in (12) and *ptdr* as in (13).

(11) *Rolala c’est l’apocalypse!! Y’a pas Ça à Gre enfin je sais pas lol on verra aux infos!*
(Oh my God, what a catastrophe! But not in Gre, well I don’t know *lol* we will see on the news!)

(12) *Ou es tu mdr me sui tt a l’arrière?*
(Where are you *mdr* are you still following?)

(13) *ta + fé tonB ta banane /la route (a) PTDR ui jri ds tr1*
(You’ve stopped dropping your banana on the road *ptdr* yes I’m laughing in the train)

- grammatical categories for *lol* such as verb, or adjective as in (1) and (2) presented above.

16. I consider interjections as a grammatical category on the same level as adjectives or verbs (see M. Haspelmath 2001).

While for the syntactic parameters I leave room for doubt, for semantic parameters, the undetermined cases are qualified as unknown, as is described in the next point.

4.2.2 *Semantic parameters*

The semantic approach to distinguishing *lol*, *mdr* and *ptdr* as discourse markers is composed of two criteria: semantic opacity and “general” meaning.

As L. Schourup (1999: 247) explains, quoting D. Wilson and D. Sperber (1993: 16), it is possible to determine whether a linguistic element has conceptual content by its semantic opacity. D. Wilson and D. Sperber (1993: 16) “claim that items which have procedural meaning are more semantically opaque than items which have conceptual content”. This assumption led me to begin the semantic analysis with semantic opacity, which I categorized into four features: literal, figurative, opaque and unknown meanings. Table 2 sets out the four meanings with a list of parameters and examples to illustrate them.

After assigning the initialisms to these four meaning categories, the occurrences are divided between conceptual, procedural and unknown meanings. To

Table 2. Parameterized and illustrated meanings

Meaning	Parameter	Example
Literal	<i>Lol</i> can be glossed by <i>éclaté de rire</i> (literally ‘burst out laughing’ or to split one’s sides with laughter), <i>mort de rire</i> (literally ‘died of laughing’ or helpless with laughter), <i>mdr</i> by <i>mort de rire</i> and <i>ptdr</i> by <i>pété de rire</i> .	(13) <i>ta + fé tonB ta banane /la route (a) PTDR ui jri ds tr1</i> (You’ve stopped dropping your banana on the road <i>ptdr</i> yes I’m laughing in the train)
Figurative	<i>Lol</i> , <i>mdr</i> and <i>ptdr</i> can be used to underline the comic or ironic aspect of a comment and so <i>lol</i> , <i>mdr</i> and <i>ptdr</i> can be glossed by <i>c’est drôle</i> ! ‘that’s funny!’, <i>c’est une blague</i> ! ‘that’s a joke!’, or as paraverbal laughter.	(14) <i>C déjà toi qui dépense t sous à sans interdit pour savoir si tu vas réussir ton permis?T’aurai du attendre la voyante lol</i> (Is it you who spends your money at Sans Interdit in order to know if you will pass your driving test? You should have waited for the psychic <i>lol</i>)
Opaque	<i>Lol</i> , <i>mdr</i> and <i>ptdr</i> have no laughter-related, comic or ironic meanings and they do not replace paraverbal laughter.	(15) <i>jte fé pl1 2 bizou partou lol.a 2m1 jtm</i> (I kiss you a lot everywhere <i>lol</i> see you tomorrow I love you)
Unknown	<i>Lol</i> , <i>mdr</i> and <i>ptdr</i> cannot be analysed since the context is insufficient for the interpretation.	(16) <i>mdr ok a ti suite!</i> (<i>Mdr</i> ok see you soon!)

distinguish conceptual meanings from procedural meanings, I chose a paradigmatic and a syntagmatic approach: the former corresponds to a set of semantic features and the latter to a relation between an introduced segment and the prior one (see B. Fraser 1999: 944). According to L. Schourup (1999: 249), this distinction in the way discourse markers convey meanings may lead to the creation of two classes of relational items. My approach here is more naïve: I consider that linguistic elements with conceptual meaning are not typical discourse markers as in (9) where *mdrr* means that the sender is laughing about her own ‘princess agenda’ joke, whereas it is procedural meaning that defines an element as a discourse marker in (17) where *Mdr* serves to continue the topic of irritation.

- (9) *Au cas si vous faites autre chose dites moi toujours je verrai pour vous caser dans mon emploi du temps de princesse... Entre 2 bals mdr bon allez a +*
(If you’re doing something else tell me and I will see if I can fit you into my princess agenda... between two balls *mdr* ok, see you soon)
- (17) *Ah ouais elle s’est enflammée comme Ça? Mdr j’té vois trop répondre, Ça devait etre tendu!!*
(So she lost her temper like that? *Mdr* I can imagine you answering back, it must have been tense!)

For this second semantic criterion, there are also occurrences with unknown meaning because the context is unavailable. The parameter used to locate this case is the lack of introduced segment, and/or prior segment as in *Exact ! Lol* (Right! *Lol*).

Doubt and ignorance have places in the syntactic and semantic criteria used in my analysis. The purpose of this tolerance is to have an inclusive approach. The next point presents the gradual approach in scoring the annotations of *lol*, *mdr* and *ptdr* to obtain a wide range of discourse markers.

4.2.3 Discourse marker scores

After the annotation of each occurrence according to the chosen syntactic and semantic parameters, I converted annotations into a score. Table 3 sums up the parameters and gives the scores assigned to the annotations. As will be noticed, the grammatical category is not taken into account in the scores because it cannot be assigned a specific score.

The chosen scores follow the literature on discourse marker characteristics. As explained in the section above, discourse markers are not attached or only loosely attached to the syntactic clause. As a consequence, the highest score is assigned to linguistic elements, which are not integrated in the syntactic clause. Concerning the position, isolation scores no points because, according to L. Schourup (1999: 231), “connectivity, however formulated (...), can be used to distinguish” discourse

Table 3. Scores of syntactic and semantic annotations

Parameter		Annotation	Score
Syntactic approach	Integration in the syntactic clause	Yes	0
		Unsure	1
		No	2
	Position	Isolated	0
		Initial, quasi-initial, medial, final (and cases of hesitation between initial and final positions)	2
Semantic approach	Semantic opacity	Literal	0
		Figurative and unknown	1
		Opaque	2
	General meaning	Conceptual	0
		Unknown	1
		Procedural	2

markers from other linguistic elements “such as illocutionary adverbials (...), attitudinal adverbials (...), and from primary interjections (...)” even if “connectivity alone is insufficient to distinguish” discourse markers “from coordinators joining intrasentential elements”.¹⁷ As for connectivity, semantic opacity is an insufficient criterion to define discourse markers but it is more typical of discourse markers. This explains why a score of zero is assigned to literal meaning and the highest score to opaque meaning. Finally, linguistic elements with conceptual meaning score zero since they are not considered as typical discourse markers, while procedural meaning scores two because it is more typical of discourse markers.

There is homogeneity in the values chosen for the discourse marker scores: the scores vary between 0 and 2 for each criterion and the score 1 is chosen for doubt and lack of knowledge (since they keep their places in this inclusive and gradual approach). Hesitation between initial and final position scores 2 because even if the specific position is unknown, it is considered as a position. Semantic lack of knowledge scores 1, like figurative meaning, because there is no evidence of opacity.

Before proceeding to the results of the discourse marker scores applied to *lol*, *mdr* and *ptdr*, and to the co-occurrences with other discourse markers as well as to the classification into moves, a word should be said about the highly divergent frequencies of the markers under investigation out of 8,000 SMSes: 429 occurrences of *lol* (5.36%), 82 of *mdr* (1.03%) and 8 of *ptdr* (0.001%) are found. These

17. Between isolation and position, there is an opposition of 0 to 2 because there is no hesitation between these two cases and the score 1 belongs to hesitation cases.

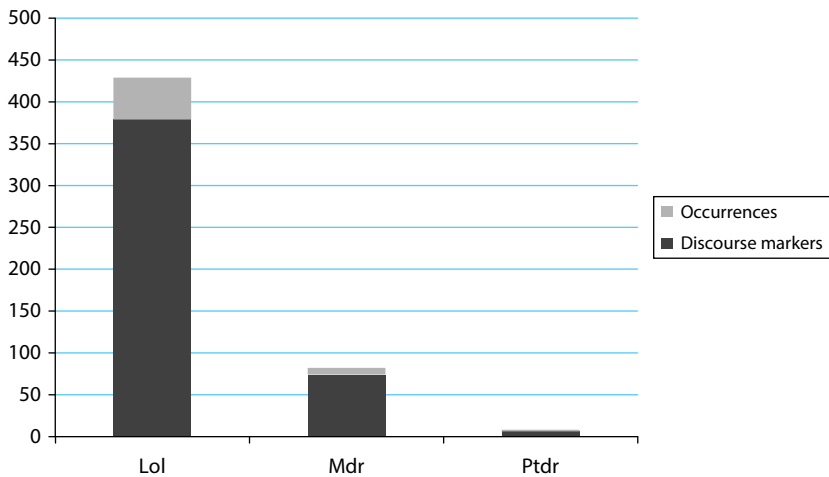


Figure 1. The discourse marker ratios relative to the total occurrences of *lol*, *mdr* and *ptdr*

results show that *lol*, *mdr* and *ptdr* handle quite differently, despite their apparent synonymy as presented in the Robert (see Section 1). In contrast, according to the syntactic and semantic characteristics of discourse markers¹⁸ in Section 2, the results of initialisms considered as discourse markers (with a score of 4 or more) are quite similar: *lol* scores 88.34% (379/429), *mdr* scores 90.24% (74/82) and *ptdr* scores 75% (6/8). The discourse marker ratios relative to the total occurrences of *lol*, *mdr* and *ptdr* are illustrated in Figure 1.

4.3 Context

After the discourse marker scores of all the occurrences of *lol*, *mdr* and *ptdr* had been calculated, 459 occurrences with a score of 4 or more were selected to be treated as discourse markers by analysing their context and relative to their moods, categorizing them in moves. These two next steps are presented in the following points.

4.3.1 Co-occurrences

From the list of linguistic elements with a discourse marker score of 4 or more, I manually checked if other discourse markers co-occur, *i.e.* they appear in contextual proximity. To avoid raising questions about the validity of the co-occurrent discourse markers, I selected only “classic” ones such as *ah* ‘ah’, *allez* ‘come on’, *alors*

18. The syntax/semantics division, calculated as discourse markers out of the overall occurrences, is indicated in Figure 2.

'then', *bah* 'well', *ben* 'well', *bon* 'well', *bref* 'in short', *donc* 'so', *du coup* 'so', *en fait* 'in fact', *en tous cas* 'anyway', *enfin* 'well', *et puis* 'and then', *mais* 'but', *oh* 'oh', *par contre* 'however', *putain* 'fuck', *quoi* 'what', *sinon* 'otherwise', *voilà* 'that's it' and reject controversial ones such as the questioned initialisms *lol*, *mdr* and *ptdr*, adverbials such as *oui* 'yes', *ok* or interjections such as *hi hi* and *ha ha*. Then I made a list of the located co-occurrences (see Appendix) to show the most frequent discourse markers in co-occurrence, namely recurrences. Although I identified some recurrences such as *lol du coup*, as in (18), I also found some discourse markers occurring with *lol*, *mdr* and *ptdr* such as *mais* as in (19), (20) and (21).

- (18) *no problemo! lol du coup pr manger je suppose que ce sera Mc do pr po faire trop cher?*
(No problem! *Lol du coup* for eating I suppose that it will be McDonalds as it's not too expensive?)
- (19) *Vivement demain lol mais tu dis à {NPERS} et {NPERS} KE JE VAIS PAS SUR LE BOOSTER !!*
(I can't wait till tomorrow *lol mais* tell {NPERS} and {NPERS} that I won't ride the bike!)
- (20) *Mdr mé c gentil tout plein ptit lou.. Toi aussi passe une bonne Nwi..*
(*Mdr mais* that's very nice young man... You too, have a good night...)
- (21) *ptdr! MAIS T OU?*
(*Ptdr! Mais* where are you?)

4.3.2 Moods and moves

The second part of the context analysis was to annotate the mood of the host sentence, namely declarative, imperative and interrogative for the occurrences of *lol*, *mdr* and *ptdr* with a discourse marker score of 4 or more. In the case of a host sentence without verbs, as in (22), or simply when the discourse marker is isolated, I annotated as *not applicable* (NA).

- (22) *Je ss o cour de ciné é je regarde «la chute de la maison d'usher»(pa le chanteur lol).*
(I'm at the film class and watching *The Fall of the house of Usher* (not the singer *lol*))

This non-applicable mood annotation makes it possible to exclude these occurrences of *lol*, *mdr* and *ptdr* from the move analysis and to focus on occurrences with sufficient discourse marker scores and annotated moods.

In order to analyse the moves, I took into account the mood of the host sentence. This move analysis follows the categorization by S. Eggins and D. Slade

(1997). There are two principal moves: opening and sustaining. The sustaining move covers two other moves: continuation and reaction. Further into the move tree, under the reaction move, there are two other moves: response and rejoinder. Table 4 sets out the functional classification of S. Eggins and D. Slade (1997) within cross-tabulations of moves to moods and references to examples to illustrate moves.

To annotate the moves, I first took into account all discourse markers within an imperative mood to see if they fit with the “request goods and services” speech act of the open move. I added to the opening move the “give advice” speech act for the imperative mood, as is the case in (23).

- (23) *Alé a la prochaine. T inkiete pa pr nisme, lol.*
(Well see you next time. Don't worry about Nisme, lol)

Next, I considered the interrogative mood where the alternative was between the opening and the rejoinder move. If, for the opening move, speech acts are more about asking for new information or requesting confirmation as in (24), for the rejoinder move, speech acts are defined as checking information as in (12) or offering an alternative position.

- (24) *bien dormi? lol t'es arrivé à l'heure?*
(Sleep well? Lol did you arrive on time?)
- (12) *Ou es tu mdr me sui tt a larière?*
(Where are you mdr are you still following?)

Finally, the declarative occurrences are classified as opening moves, when the associated speech act is about giving information as in (25), as continuing moves for clarifying as in (26), as responding moves for showing willingness to respond and ability to provide requested information as in (11) and as rejoinder moves for confirmation, agreement, disagreement, contradiction as in (27).

- (25) *Sinn rien de 9 g un nouveau g lol et toi?*
(Otherwise nothing new I have a new mobile lol and you?)
- (26) *vous pouvez rpd part email si vous voule coute moind cher lol*
(You can reply by e-mail if you want to spend less lol)
- (11) *Rolala c'est l'apocalypse!! Y'a pas Ça à Gre enfin je sais pas lol on verra aux infos!*
(Oh my God, what a catastrophe! But not in Gre, well I don't know lol we will see on the news!)
- (27) *Lol c bien ske jdit c du chantage!*
(Lol it's what I say it's blackmail!)

Table 4. Speech functions within cross-tabulations of moves to moods illustrated

	Open ¹⁹	Sustain		
		Continue	React	
			Respond ²⁰	Rejoinder ²¹
Declarative	Give factual/attitudinal/evaluative information (see example 25)	Clarify, exemplify or restate; offer additional contrasting information; qualify previous move by giving details of time, place, cause, conditions, etc. (see examples 8 and 26)	Indicate inability to provide requested information (see example 11)	Volunteer further details/implications for confirmation; provide clarification, acquiesce with information; dismiss addressee's right to his/her position; contradict importance of a challenge (see example 27)
	Request goods and services (see example 4 and 23)			
Imperative				
	Give goods and services; request (confirmation/agreement with) factual information/ (agreement with) subjective information (see example 24)			Elicit repetition of a misheard element or move; verify information heard; get additional information needed to understand prior move; question relevance, legitimacy, veracity of prior move; offer alternative position (see example 12)

19. Unspecified moods include "attention seeking" as an opening move.

20. Unspecified moods include "show willingness to interact by responding to salutation, etc.", "display attention to the speaker", "carry out request for goods and services", "accept proffered goods and services", "indicate support for information given" or "knowledge of information given", "provide information requested", or positive/negative "response to question", "indicate inability to comply with prior command", "deny acknowledgment of information", and "negate prior information" as a responding move.

21. Unspecified moods include "terminate interaction" as a rejoining move.

5. Results

The discourse marker results given at the end of Section 4.2 (see Figure 1) need to be detailed with the ratios of syntactic and semantic parameters taken into account to obtain discourse marker scores, namely integration in syntactic clause, position, semantic opacity and procedural meaning. *Lol* appears with 97.67% (419/429) non-integrated uses, 56.88% (244/429) positioned uses, 82.1% (352/429) opaque uses, and 55.94% (240/429) procedural uses. *Mdr* has 97.56% (80/82) non-integrated uses, 64.63% (53/82) positioned uses, 28.05% (23/82) opaque uses, and 17.07% (14/82) procedural uses. Out of a total of 8 occurrences, *ptdr* shows non-integrated uses for every occurrence, positioned uses for three occurrences, an opaque use in one instance, and a procedural use in another instance. Figure 2 gives an overview of the syntactic and semantic distribution of the three discourse markers.

It is clear that almost all initialisms are not integrated in the syntactic clause. Concerning their position, *ptdr* is preferred in isolation while *lol* and *mdr* are mostly positioned. In contrast, semantic parameters show high divergence in the sense that *lol* is the most semantically opaque and conveys the most procedural meaning while *mdr* and *ptdr* show very low opacity and procedural meaning occurrences.

Although *lol* and *mdr* have more discourse marker uses than *ptdr*, they have lower co-occurrence ratios: *lol* co-occurs in 8.71% of cases (33/379), *mdr* in 8.11% (6/74), and *ptdr* in one in six occurrences (16.7%). These co-occurrences are listed with the frequency of recurrences in Appendix.

With respect to mood, the results for the moves of the discourse marker host sentences are quite similar: *lol* is found in 35% of cases in opening moves, 31% in continuing moves, 13% in responding moves and 21% in rejoinder moves; *mdr* is found in 37% of cases in opening moves, 33% in continuing moves, 5% in responding moves and 26% in rejoinder moves; *ptdr* occurs in 33% of cases in opening moves, 33% in continuing moves, and 33% in responding moves. The move distribution of these three discourse markers is detailed in Table 5.

These move distribution results are quite similar if the responding and rejoinder moves are taken together as the reacting move. The reacting move accounts for about 34.23% of occurrences of *lol*, 30.23% of occurrences of *mdr* and 33.33% of occurrences of *ptdr*. In short, all three moves account for between 30 and 35% of occurrences; thus none of the three initialisms shows a trend to coincide with a particular move, but instead there is a balanced spread of speech act uses. Nevertheless, these results need to be discussed.

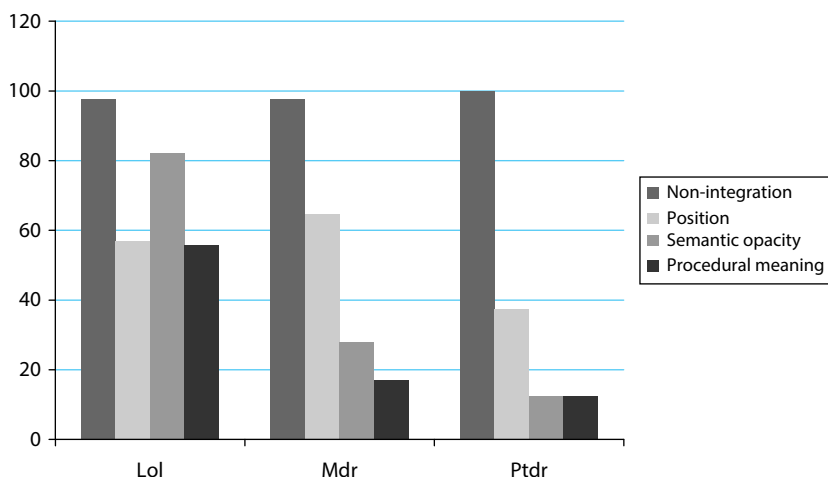


Figure 2. Syntactic and semantic distributions of *lol*, *mdr* and *ptdr*

Table 5. Move distribution for *lol*, *mdr* and *ptdr*

Discourse marker	Open	Continue	React	
			Respond	Rejoinder
<i>Lol</i>	34.76% (65/187)	31.02% (58/187)	13.37% (25/187)	20.86% (39/187)
<i>Mdr</i>	37.21% (16/43)	32.56% (14/43)	4.65% (2/43)	25.58% (11/43)
<i>Ptdr</i>	33.33% (1/3)	33.33% (1/3)	33.33% (1/3)	0

Discussion

In comparison with *mdr* and *ptdr*, *lol* is used more than five times more frequently and more than 53 times more frequently in the corpus of 8,000 SMSes. This means, first, that the result presentation expressed in percentage terms is not representative of the reality, although it does ensure readability, second, that the consecutive syntactic and semantic distribution, discourse marker scores, co-occurrence list and move distribution should be considered in the light of the uneven occurrence of *lol*, *mdr* and *ptdr*, and third, that it is quite difficult to generalize the results for *ptdr* with only eight occurrences.

In addition, the number of discourse marker uses of *lol*, *mdr* and *ptdr* is probably an incorrect estimate, since no access is given through the corpus to the complete exchange of SMSes between sender and recipient. This leads to a lot of unknown cases and hesitations that could overestimate the discourse marker scores by giving an additional point, since the approach of this work is inclusive and gradual.

Again, the position analysis could be controversial because it is based on punctuation to simplify the analysis. However, it is known that punctuation varies from one writer to another, especially in texting where creativity also extends to orthotypography. This question also raises the issue of syntactic scope. It is not certain that the isolated position has no syntactic scope, and the analysis in moves underlines this issue. In the example (28), it is not certain whether *mdr* is hosted by the previous sentence, which finishes with a full stop, or by the next one, which finishes with a winking smiley.

- (28) *mais pourquoi tu m'as pas envoyé ce sms pour la science.mdr, mais t'as bien fais ;)*
 (but why didn't you send me this sms for science. *Mdr*, but you did well ;))

There could be also a semantic scope issue, since the boundaries between syntax and semantics seem unclear. Syntactic and semantic scope should be taken into consideration in a future study.

Conclusions

The corpus analysis has shown that *lol* is used more often than *mdr* and *ptdr* in texting and that these three initialisms act in quite a similar way. The results confirm that *lol*, *mdr* and *ptdr* could work as discourse markers syntactically and semantically in text messages.

Considering each discourse marker separately, *lol* is the most semantically opaque and conveys the most procedural meaning, while *mdr* and *ptdr* show very low opacity and procedural meaning. This raises the question of the semantic specificity of the most frequent discourse marker initialism, *lol*, and the possibility that this meaning is specific to text messaging communication.

As discourse markers, all three initialisms co-occur with other discourse markers and they seem not to have been linked to specific speech acts, since their host sentence is equally likely to be an opening, continuing or reacting move.

In texting, the common initialisms *lol*, *mdr* and *ptdr* can be used as discourse markers and have a wide range of speech act features. Nevertheless, there are a lot of unknown cases and hesitations because of the constraints of the SMS corpus and of SMS language in itself. These cases raise the question of syntactic and semantic scope, which needs to be taken into consideration in further research. A corollary question is whether certain discourse markers tend to occur in a type of computer-mediated communication. Unravelling this specific role must, however, remain the topic of ongoing and future research through the comparison of these results with results from chat and oral corpora.

Acknowledgements

I would like to thank Liesbeth Degand, my PhD supervisor, who helped me in the preparation of this paper. My special thanks are extended to Catherine Bolly and Federica Ciabbarri for interesting discussions on discourse marker defining parameters. I am particularly grateful to Louise-Amélie Cougnon for giving me information about the sms4science project.

This project is financially supported by a Special Research Grant from the Université catholique de Louvain (Louvain-La-Neuve).

References

- Aijmer, Karin; Anne-Marie Simon-Vandenberg. 2009. Pragmatic Markers. In *Handbook of Pragmatics*, J.-O. Östman, J. Verschueren (eds.), 13th Annual Instalment, Amsterdam: John Benjamins. Available on www.benjamins.nl/online
- Baron, Naomi. 2004. See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23: 397-423.
- Bazzanella, Carla. 2006. Discourse Markers in Italian: Towards a "Compositional" Meaning. In *Approaches to Discourse Particles*, K. Fischer (ed.), 449-464, Amsterdam: Elsevier.
- Benzitoun, Christophe; Anne Dister; Kim Gerdes; Sylvain Kahane; Paola Pietrandrea; Frédéric Sabio. 2010. *Tu veux couper là faut dire pourquoi*: Propositions pour une segmentation syntaxique du français parlé. *CMLF 2010, La Nouvelle Orléans*, 16 p. Available on: http://www.linguistiquefrancaise.org/index.php?option=com_article&access=standard&Itemid=129&url=/articles/cmlf/abs/2010/01/cmlf2010_000201/cmlf2010_000201.html
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Blanche-Benveniste, Claire; Mireille Bilger; Christine Rouget; Karel Van den Eynde. 1991. *Le français parlé. Etudes grammaticales*. Paris: Editions du CNRS.
- Blanche-Benveniste, Claire; José Deulofeu; Jean Stéfanini; Karel Van den Eynde. 1984. *Pronom et syntaxe. L'approche pronominale et son application au français*. Paris: Société d'études linguistiques et anthropologiques de France.
- Chafe, Wallace L. 1982. Integration and Involvement in Speaking, Writing, and Oral Literature. In *Spoken and Written Language: Exploring Orality and Literacy*, D. Tannen (ed.), 35-53, Norwood, New Jersey: Ablex Publishing Corporation.
- Chafe, Wallace; Jane Danielewicz. 1987. Properties of Spoken and Written Language. In *Comprehending Oral and Written Language*, R. Horowitz, S.J. Samuels (eds.), 83-113, San Diego: Academic Press.
- Cougnon Louise-Amélie; Gudrun Ledegen. 2008. « c'est écrire comme je parle » Une étude comparative des variétés du français dans l'écrit sms. In *Les Voix des Français*, M. Abecassis, G. Ledegen (eds.), Vol. 2, *Modern French Identities*, 94, 39-57, Bern: Peter Lang.
- Eggs, Suzanne; Diana Slade. 1997. *Analysing Casual Conversation*. London, Oakville: Equinox Publishing Ltd.
- Fairon, Cédric; Jean Klein; Sébastien Paumier. 2006. Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation [CD-Rom]. *Cahiers du Cental* 3.2, Louvain-La-Neuve: Presses universitaires de Louvain.

- Fischer, Kerstin. 2006. Towards an Understanding of the Spectrum of Approaches to Discourse Particles: Introduction to the Volume. In *Approaches to Discourse Particles*, K. Fischer (ed.), 1-20, Amsterdam: Elsevier Ltd.
- Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931-952.
- Fraser, Bruce. 2006. Towards a Theory of Discourse Markers. In *Approaches to Discourse Particles*, K. Fischer (ed.), 189-204, Amsterdam: Elsevier.
- Hård af Segerstad, Ylva. 2005. Language in SMS: A Socio-Linguistic View. In *The Inside Text: Social, Cultural and Design Perspectives on SMS*, R. Harper, L.A. Palen, A.S. Taylor (eds.), 33-51, Dordrecht: Kluwer Academic Publishers.
- Haspelmath, Martin. 2001. Word Classes and Parts of Speech. In *International Encyclopedia of the Social and Behavioral Sciences*, P.B. Baltes, N.J. Smelser (eds.), 16538-16545, Amsterdam: Pergamon.
- Herring, Susan C. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4.4. Available on <http://jcmc.indiana.edu/vol4/issue4/herring.html>
- Jakobson, Roman. 1960. Linguistics and Poetics. In *Style in Language*, T.A. Sebeok (ed.), 350-377, Cambridge, MA: MIT Press.
- Jansen, Erin. 2003. *Netlingo: The Internet Dictionary*. Ojai, CA: Netlingo, Inc.
- Koch, Peter; Wulf Oesterreicher. 2001. Gesprochene Sprache und geschriebene Sprache. Langage parlé et langage écrit. In *Lexicon romanistischen Linguistik*, G. Holtus, M. Metzeltin, C. Schmitt (eds.), 1.2: 584-627, Tübingen: Max Niemeyer.
- Lewis, Cynthia; Bettina Fabos. 2005. Instant messaging, literacies, and social identities. *Reading Research Quarterly* 40: 470-501.
- Lewis, Diana. 2006. Discourse Markers in English: A Discourse-Pragmatic View. In *Approaches to Discourse Particles*, K. Fischer (ed.), 43-59, Amsterdam: Elsevier.
- McCracken, James (ed.). 2012. *The Oxford English Dictionary. The Definitive Record of the English Language*. Online version. Oxford: Oxford University Press. Available on: <http://www.oed.com/>
- Merchant, Guy. 2001. Teenagers in cyberspace: An investigation of language use and language change in internet chatrooms. *Journal of Research in Reading* 24: 293-306.
- Michot, Nicolas. 2007. Les usages lexicaux des jeunes sur les supports modernes de communication. *Acte du 26e Colloque international Lexique Grammaire, Bonifacio, France, 2-6 octobre 2007*, 8 p. Available on: <http://infoling.univ-mlv.fr/Colloques/Bonifacio/proceedings/michot.pdf>
- Panckhurst, Rachel. 2009. Short Message Service (SMS): typologie et problématiques futures. In *Polyphonies, pour Michelle Lanvin*, T. Arnavielle (coord.), Université Paul-Valéry Montpellier 3: 33-52. Available on: <http://hal.archives-ouvertes.fr/hal-00443014>
- Robert, Paul. 2013. *Le Petit Robert. Dictionnaire alphabétique et analogique de la langue française*. J. Rey-Debove, A. Rey (dir). Paris: Le Robert.
- Robert, Paul. 2011. *Le Petit Robert de la langue française numérique* (CD-Rom). J. Rey-Debove, A. Rey (dir). Paris: Le Robert.
- Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge/New York/Melbourne: Cambridge University Press.
- Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107: 227-265.
- Spatafola, Julia Nadine. 2008. *IM learning 2 write? A study on how instant messaging shapes student writing*. Unpublished Master's thesis, Queens University, Kingston, Ontario, Canada.
- Tagliamonte, Sali A.; Derek Denis. 2006. *LOL for Real! Instant Messaging in Toronto Teens*. Toronto: Linguistic Association of Canada and the United States (LACUS).

- Tagliamonte, Sali A.; Derek Denis. 2008. Linguistic ruin? Lol! Instant messaging and teen language. *American Speech* 83(1): 3-34.
- Varnhagen, Connie K.; G. Peggy McFall; Nicole Pugh; Lisa Routledge; Heather Sumida-MacDonald; Trudy E. Kwong. 2010. Lol: New language and spelling in instant messaging. *Reading and Writing* 23: 719-733.
- Vincent, Diane. 2005. The journey of non-standard discourse markers in Quebec French: Networks based on exemplification. *Journal of Historical Pragmatics* 6/2: 188-210.
- Werry, Christopher C. 1996. *Linguistic and Interactional Features of Internet Relay Chat*. Philadelphia, PA: John Benjamins Publishing.
- Wilson, Deirdre; Dan Sperber. 1993. Linguistic form and relevance. *Lingua* 90: 1-25.

Appendix. List of co-occurrences and recurrences with *lol*, *mdr* and *ptdr*

Discourse marker	Co-occurrence	Recurrence (number of occurrences)
Lol	<i>Ah lol</i> 'ah lol'	<i>Ah lol</i> (2)
	<i>Lol ah ben</i> 'lol ah well'	(1)
	<i>Lol allez</i> 'lol come on'	<i>Lol allez</i> (2)
	<i>Lol alors</i> 'lol then'	(1)
	<i>Alors lol ben</i> 'then lol well'	(1)
	<i>Lol bah</i> 'lol well'	(1)
	<i>Lol ben</i> 'lol well'	<i>Lol ben</i> (3)
	<i>Lol bon allez</i> 'lol well come on'	(1)
	<i>Lol bref</i> 'lol in short'	(1)
	<i>Lol donc</i> 'lol so'	(1)
	<i>Lol du coup</i> 'lol so'	(1)
	<i>Lol en fait</i> 'lol in fact'	(1)
	<i>Lol en tous cas</i> 'lol anyway'	<i>Lol en tous cas</i> (2)
	<i>Lol enfin</i> 'lol I mean'	(1)
	<i>Lol et puis</i> 'lol then'	(1)
	<i>Lol fin bon</i> 'lol I mean, well'	(1)
	<i>Lol mais</i> 'lol but'	<i>Lol mais</i> (4)
	<i>Lol mais en fait</i> 'lol but in fact'	(1)
	<i>Lol oh</i> 'lol oh'	(1)
	<i>Lol par contre</i> 'lol but'	(1)
	<i>Quoi lol</i> 'what lol'	(1)
	<i>Sinon lol</i> 'otherwise lol'	<i>Sinon lol</i> (2)
Mdr	<i>Mdr bon allez</i> 'mdr well come on'	(1)
	<i>Mdr bon bah voilà alors</i> 'mdr well then that's it then'	(1)
	<i>Mdr mais</i> 'mdr but'	<i>Mdr mais</i> (2)
	<i>Mdr putain</i> 'mdr fuck'	(1)
	<i>Putain mdr</i> 'fuck mdr'	(1)
Ptdr	<i>Ptdr mais</i> 'ptdr but'	(1)

Summary

Given the extensive use of *LOL* (*Laughing Out Loud*), *MDR* (*Mort De Rire*) and *PTDR* (*PéTé De Rire*) in French texting and the inclusion of *lol* in the Oxford English Dictionary in 2011 and in Le Petit Robert in 2013, this paper aims to study the functioning of these three initialisms. Considered mainly as interjections, I hypothesize that *lol*, *mdr*, and *ptdr* could work syntactically (through their non-integration in the syntactic clause and position) and semantically (through their semantic opacity and procedural meaning) as discourse markers. In order to show this discourse marker use, this exploratory study was carried out using a prototypical methodology characterized by an inclusive and gradual approach. Based on syntactic and semantic analysis, a discourse marker ranking was performed, and this resulted in a contextual analysis (co-occurrences, moods and moves).

This inclusive and gradual approach allows for doubt and lack of knowledge of the context in the analysis process. More broadly, this work also throws light on texting as a spontaneous computer-mediated communication type with writing constraints imposed by the communication medium and the situation.

Index

A

abbreviations 2, 4, 5, 30, 43,
44, 48n1, 55, 64, 73, 74, 75,
75n2, 76, 80, 81, 83, 84, 92,
94, 96, 97, 98, 99, 103, 104,
106, 113, 116n2, 122, 124,
146, 146n12, 154, 171, 219,
220, 229, 231, 232, 239, 240
accent stylisation 73, 81, 83
acronyms 30, 55, 73, 74, 75,
75n2, 79, 80, 81, 83, 84, 85,
171, 239, 240, 241, 242
adolescents 4n2, 8, 9, 29, 30,
31, 32, 33, 41, 43, 45, 47,
111, 124, 137, 161, 164,
188, 222, 225
alternation 112, 116, 119,
128, 129, 130, 131, 134
annotation 7, 89, 90, 91, 92,
93, 94, 95, 96, 97, 98, 99,
103, 104, 106, 107, 107n20,
108, 155, 223, 251, 252, 254
annotation score 251
anonymisation 8, 12, 13, 14,
15, 18, 19, 20, 22, 23, 26,
26n9, 27, 28, 91, 94, 141n1,
153, 153n17, 154, 156,
160, 164
anonymisation software 153

B

borrowing 112, 120, 121,
129, 130, 138

C

clippings 30, 38, 73, 74, 75,
77, 80, 81, 83, 229, 230,
231, 234
clitic subjects 200, 203, 209,
212, 213, 215
code-switching 8, 99, 103,
106, 111, 112, 115, 117,

120, 122, 128, 130, 135,
136, 138, 139, 152, 182, 186
cognitive cost 49, 51, 54, 61,
62, 65
community of practice 112,
113, 122, 123, 127, 130,
134, 135
conceptual meaning 244 ,
250, 251, 252
contractions 30, 68, 69, 73,
74, 80, 81, 83, 181, 229,
231, 234, 237
contributor's profile 99
controlled processes 49

D

diachronic analysis 68, 71, 85
dialect 1, 123, 124, 135, 149,
221, 229, 231, 232, 233,
234
DICER 217, 218, 223, 225,
226, 227, 227n10, 228, 229,
230, 233, 234, 235, 237
discourse analysis 10, 85,
190, 215, 220n2, 237, 240
discourse marker 251, 258,
262
dual task 56, 57, 59

E

eSMS 48, 48n1, 50, 51, 54,
61, 62, 65, 163
expertise 56, 58

F

formulae 131, 135
French 11, 13, 15, 17, 24, 28,
29, 33, 34, 35, 36, 37, 38,
41, 43, 50, 64, 87, 88, 89,
91, 92, 93, 96, 97, 98, 99,
100, 102, 105, 106, 107n20,

108, 109, 114, 115, 116,
116n2, 117, 119n6, 120,
121, 122, 124, 125, 126,
126n17, 126n18, 127n19,
128, 129, 129n23, 130,
134n36, 136, 138, 141, 144,
145, 146, 147, 149, 150,
152, 152n15, 154, 154n18,
156, 157, 161, 162, 163,
164, 186, 191, 192, 193,
194, 195, 195n1, 196, 197,
198, 199, 200, 201, 202,
203, 205, 206, 207, 208,
209, 210, 211, 212, 213,
214, 215, 217, 221, 222,
235, 236, 237, 240, 241,
242, 242n4, 242n5, 246,
249, 260, 262, 263

French language 15, 36, 88

G

g-clippings 30, 38, 73, 75, 77,
80, 81, 83
gender 260
German 24, 36, 84, 103,
114, 115, 116, 116n2, 119,
119n4, 119n5, 119n6, 120,
121, 121n11, 121n7, 121n8,
122, 123, 123n12, 123n13,
123n14, 124, 125, 125n15,
125n16, 126, 126n17,
126n18, 127n19, 127n20,
128, 128n21, 129, 129n22,
129n23, 130, 130n24,
131n27, 132n29, 132n30,
132n31, 133n32, 133n33,
133n34, 133n35, 134,
134n36, 137, 138, 152n15,
186, 191, 200, 235
grassroots literacy 169, 173,
175

- H**
heteroglossia 186
heterography 175
heuristic algorithms 25
hybrid language use 111,
122, 126
- I**
imperative 197, 198, 199, 201,
206, 207, 208, 209, 210,
211, 212, 213, 254, 255, 256
initialisms 30, 68, 73, 74, 75,
75n2, 79, 80, 81, 83, 240,
240n2, 241, 242, 243, 246,
247, 250, 253, 254, 257,
259, 263
insertion 112, 116, 127, 128,
130, 131, 134, 156, 226,
226n8, 227
intonation 82, 83, 160, 220
Italian 36, 93, 114, 116,
116n2, 118, 121, 121n11,
124, 134n36, 152, 152n15,
200, 260
- L**
literacy 43, 44, 64, 74, 169,
170, 171, 173, 175, 176,
182, 184, 185, 186, 188,
218, 219, 221, 222, 236, 260
logograms 74
longitudinal study 32, 33
- M**
mediated electronic discourse
160
misspellings 30, 73, 75, 80,
81, 83, 230, 232, 236
mixed spellings 117, 124,
126, 128, 135, 139
mood 254
- N**
Natural Language Processing
13, 17
negation marking 191
network ethnography 188
non-conventional spelling
31, 73, 76, 81, 82, 83, 178
normalisation 88, 90, 91, 92,
217, 218, 222, 223, 224,
225, 226, 226n7, 226n9,
227, 228, 232, 234, 235, 237
- O**
orthographic forms 30
- P**
pictograms 74
plurilingual communication
111, 113
pragmatics 8, 43, 44, 136,
138, 161, 215, 260, 261,
262
procedural meaning 240,
244, 250, 251, 252, 257,
258, 259, 263
pronunciation 31, 36, 68,
73, 74, 75n2, 76, 79, 80,
81, 82, 83, 85, 116n2, 122,
126, 222
pseudo-borrowings 117,
120, 122
- R**
reductions 50, 55, 56, 57, 58,
59, 61, 74, 78, 79, 80, 81,
83, 132, 232
register 5, 30, 42, 45, 97, 164,
180, 181, 194, 205, 212, 220
respellings 74, 75, 75n3, 76,
78, 79, 81, 83, 84, 219, 220,
221, 229, 230, 233
- S**
Seek&Hide 8, 11, 12, 13, 14,
15, 16, 18, 20, 21, 22, 23,
24, 25, 26, 26n9, 28, 153,
161
semantic opacity 252
shortenings 30, 73, 74, 80,
81, 83
smileys 29, 55, 73, 74, 76, 78,
83, 93, 94, 96, 98, 104, 116,
155, 156, 156n21, 159, 242,
245, 248
sms abbreviations 122, 154
sms normalisation 88
SMS writing 13, 15, 24, 26,
48, 53, 61, 65, 147, 149,
150, 151n14, 159
sms4science 6, 6n4, 7, 11,
11n2, 27, 29, 41, 44, 52, 88,
89, 90, 95, 109, 114, 124,
130, 137, 139, 141n1, 144,
144n7, 144n9, 150, 155,
161, 162, 164, 192, 194,
200, 213, 214, 215, 246,
246n8, 260
social agency 169, 173, 184
social capital 170
sociolinguistic analysis 99,
169, 172
sociolinguistic questionnaire
7, 142, 146, 147, 149, 151
sociolinguistic variable 194,
213
sound-to-spelling relationship
72, 79, 83
speech function 241, 247,
256
spelling variants 124, 217,
219, 221, 223, 225, 226,
227, 228, 229, 230, 233,
234, 237
spelling 43, 51, 52, 52n2,
52n3, 55, 95, 137, 186, 213,
217, 219, 220, 221, 232,
234, 235, 236, 237
spelling adaptations 67, 73,
74, 77
spelling trends 217, 224, 226,
227, 234, 235, 237
spelling variation 217, 218,
219, 220, 221, 222, 223,
225, 226, 227, 228, 230,
232, 233, 234, 235, 237
style shifting 117
subject drop 201, 204, 205,
207, 211, 212, 215
Swiss French 36, 116, 199,
209, 212
Swiss German 114, 115, 116,
116n2, 119n6, 120, 121,
121n11, 121n8, 123, 124,
125, 126, 128, 129, 130,
132n31, 134, 134n36, 200
syntactic integration 247,
249, 257, 263
syntactic position 247, 248,
249, 257, 259
- T**
tagging 23, 122, 123, 223,
237
Technoliteracy 170n2
Textisms 2, 37, 38
texto4science 87, 89, 91, 92,
99, 105, 105n17, 106, 108,
109, 144n9, 162

tokenisation 15
transcoding 141n1, 146, 154,
160, 164
transidiomaticity 176, 180,
183, 184
transnational migration 182
type of keyboard 103, 104

V
VARD 223, 225, 234, 235

W
working memory 48, 63, 64
written language 192, 193,
198

X
xenoglossy 176, 179, 180,
183, 185
XML schema 90, 91