Collecting a Corpus of Dutch SMS

Maaske Treurniet[1], Orphée De Clercq[2,3], Henk van den Heuvel[1] and Nelleke Oostdijk[1]

CLST, Centre for Language and Speech Technology, Radboud University Nijmegen[1]
Erasmusplein 1, 6525 HT Nijmegen The Netherlands
{m.treurniet, h.vandenheuvel, n.oostdijk}@let.ru.nl
LT3, Language and Translation Technology Team, University College Ghent[2]
Groot-Brittanniëlaan 45, 9000 Ghent Belgium
orphee.declercq@hogent.be
Dept. of Applied Mathematics and Computer Science, Ghent University[3]
Krijgslaan 281 (S9), 9000 Gent Belgium

Abstract

In this paper we present the first freely available corpus of Dutch text messages containing data originating from the Netherlands and Flanders. This corpus has been collected in the framework of the SoNaR project and constitutes a viable part of this 500-million-word corpus. About 53,000 text messages were collected on a large scale, based on voluntary donations. These messages will be distributed as such. In this paper we focus on the data collection processes involved and after studying the effect of media coverage we show that especially free publicity in newspapers and on social media networks results in more contributions. All SMS are provided with metadata information. Looking at the composition of the corpus, it becomes visible that a small number of people have contributed a large amount of data, in total 272 people have contributed to the corpus during three months. The number of women contributing to the corpus is larger than the number of men, but male contributors submitted larger amounts of data. This corpus will be of paramount importance for sociolinguistic research and normalisation studies.

Keywords: SoNaR, text messaging, corpus collection

1. Introduction

Within he Flemish-Dutch SoNaR project a 500-million-word corpus[1] of written Dutch has been built. This corpus was designed to serve as a general reference on all kinds of research on language and language usage (Oostdijk et al., 2008). One of the main novelties is that it explicitly aimed to include, besides the more traditional text genres, a large variety of digital media such as chats, tweets (Sanders, 2012) and text messages (SMS).

In this paper we introduce a corpus of Dutch SMS (Short Message Service), which has been collected in the Netherlands and Flanders over a time span of about three months. The SoNaR SMS corpus[2] contains about 53K text messages representing a one-third – two-thirds Dutch/Belgian spread. This corpus is, to the best of our knowledge, the first freely available Dutch SMS collection.

Ensuring free availability was one of the main prerequisites of the SoNaR project. This presumes that all Intellectual Property Rights (IPR) are cleared to the fullest possible extent, turning the actual data collection process into a challenging task. Previous successful data collection techniques (such as described in De Clercq and Montero Perez (2010)) could not be followed. Besides, various technical characteristics are involved that further complicate this process. The length of text messages is restricted to 140 bytes or 160 seven-bit characters. Sending SMS is a paying service and typing the actual text requires using small keys on a mobile phone. These three factors have led to the creation of a contested language variant called "texting" (Crystal, 2008).

Inspired by previous successful SMS collection projects, we decided to collect SMS texts on a large scale based on voluntary contributions. In order to reach a broad audience we employed the regional and national media in both countries. The effect of media coverage is investigated in closer detail throughout the paper and we show that especially free publicity in newspapers and on social media networks results in more contributions.

Besides format conversion, anonymisation and performing some basic tokenization, the gathered text messages have not been further processed because the SoNaR project only aimed to collect them. Based on the collection itself, however, we were able to draw some interesting findings on SMS usage and gender characteristics in the Low Countries. Though men seem more likely to use smartphones, women are more willing to contribute their text messages to a corpus.

In the remainder of this paper we first describe other SMS collection projects after which the SoNaR SMS corpus is introduced. We continue by discussing the influence of media coverage and by revealing some interesting tendencies in section 4. We finish with concluding remarks and prospects for future work.

2. Related Work

Since the first SMS service was offered to consumers in 1993 it has become one of the most widespread means of

[1] For more information we refer to Oostdijk et al. (forthcoming).

[2] The corpus will be distributed by the Dutch-Flemish HLT Agency (TST Centrale): http://www.inl.nl/tst-centrale/ as part of the SoNaR corpus.

communication, especially among youngsters. In 2010 alone, about 6.1 trillion text messages were sent worldwide.[3] On average, a person sends around 25 messages a month, whereas the average American teenager texts about 80 times a day.
This widespread usage has drawn the attention of many researchers from different strands. In Tagg (2009) an overview is presented of SMS-related research focuses. Among others, the focus has been on analyzing conversational 'threads' and abbreviations, determining how written communication adapts to technology (Grinter & Eldridge, 2001), on conducting social-scientific studies into the communicative practices of mobile technology (Kasesniemi & Rautianen, 2002), on improving predicted text entry (How, 2004) or on sociolinguistic research (Grant, 2009).

From these various research opportunities one can easily deduct which metadata users might require. For sociolinguistic purposes, background information concerning the author of each SMS (age, gender, city, country of residence) is required. Moreover, a message's time and date can be helpful in studying the behavior of SMS communication during various moments of the day or for diachronic studies. Besides these metadata, an exact transcription of the text (including typing errors, smileys and abbreviations) is needed for linguistic studies. When it comes to improving the existing technology such as predictive text entry or developing new text entry methods, it is useful to have metadata along with the corpus, related to the type of mobile device and the texting habits.
In current Natural Language Processing research, SMS data are at the heart of normalization studies (Beaufort et al., 2010). Normalization of noisy data becomes a big challenge since state of the art text processing tools (tokenisers, taggers, chunkers) have been trained on 'clean' text and fail when applied to user generated content.[4] For machine learning and other purposes it is useful to know the number of messages contributed by one author and the distribution of this number among the contributors.

What is lacking, however, are freely available data sets in which this information is included and on which these types of research can be conducted. In general, SMS corpora are scarce and the data are often not publicly available (Chen & Kay, forthcoming). This is mostly because of the private character of SMS. The same is valid for Dutch; there is currently no freely available Dutch SMS collection.
Existing SMS corpora differ in size, language and collection method. Two notable SMS collection projects are the sms4science project[5] and the NUS SMS Corpus Project[6]. Sms4science was started up in Belgium and over the years the same techniques have been carried out in other countries (Switzerland, France, Greece, Spain and Italy). The NUS team on the other hand, focused on collecting English and Mandarin text messages.
Looking at these and other SMS collection projects, basically three different collection methods can be distinguished (Chen & Kay (forthcoming)). The first method can be characterized as 'the recruitment of acquaintances'. Here, personal and/or professional contacts are used to collect SMS messages. This method was followed by Bieswanger (2006) for the collection of the English Language corpus and the German SMS corpus.
A second method is described in Herring & Zelenhauskaite (2009). They present an Italian SMS corpus comprising SMS messages from iTV SMS. This service enables viewers to send SMS directly to the channel, which are then briefly displayed at the bottom of the screen. All text messages included in the corpus are thus original, i.e. no adjustments were made to the text itself, and they originate from the same source.
A third method of SMS collection concerns large scale corpus building by employing technical means, i.e. to extract or copy the text from SMS messages directly from a device. In this way potentially large amounts of data can be obtained once the technical support is created. For the actual collection, however, collaboration is still required from either the phone companies or phone owners. Phone companies have very restricted legal regulations whereas persuading phone owners is time-consuming and subject to some ethical considerations. The sms4science project is a good example of the first approach and the NUS SMS Corpus of the latter. During sms4science the barrier for donation was lowered by letting people forward their messages directly to a central number free of charge (Fairon & Faumier, 2006). The NUS team on the other hand developed an application on the Google Android platform that allowed users to automatically send messages to the corpus (Chen & Kay, forthcoming).

3. SoNaR SMS Corpus

For the collection of SMS within the framework of the SoNaR project a combination of the above-mentioned data collection methods was employed. Our main objective was to obtain large data quantities from various user groups. At the same time, however, it had to be easy for the contributors to donate data while ensuring privacy.
Because of SoNaR's strict IPR requirements it was decided to directly contact the phone owners and only include 'sent' SMS messages in the corpus. Only then can a user be considered the actual owner of a message and this approach also enabled us to collect a substantial amount of metadata.
For the actual donation we translated the NUS Android application to Dutch and modified it to our purposes, i.e. the manual was updated with the conditions for contribution to SoNaR.
Although the Android application met our needs best, it was decided to also include alternative ways for contributing SMS. For this purpose a project website[7] was set up containing instructions on how various users could donate text messages.

[3] http://www.itu.int/ITUD/ict/material/ FactsFigures2010.pdf
[4] During LREC 2012 a workshop is devoted to this subject: NLP4UCG.
[5] http://www.sms4science.org/?q=en
[6] http://wing.comp.nus.edu.sg:8080/SMSCorpus/
[7] www.sonarproject.nl or www.sonarproject.be (in Dutch)

- Smartphone users, using the Android platform, could download an application that automatically uploads all sent SMS messages to their Gmail mailbox account. Afterwards, this list could be sent to the SoNaR SMS project;
- Apple iPhone[8] and Nokia users could find instructions on the project website on how to obtain the SMS back-up file when connecting their phone to a computer;
- All other mobile phone users could fill in an online submission form and manually retype some text messages.

3.1 Anonymisation and Metadata

A main consideration in creating an SMS corpus is the need to protect the rights and interests of both the authors and other persons mentioned in the text messages, while still preserving the original text and gathering sufficient metadata information.
In order to protect the identity of each contributor, phone numbers have been encrypted inside the corpus. They have all been replaced with a unique identifier, so that the end-user is still able to locate multiple messages coming from the same contributor. Besides phone numbers, other private data inside the messages have also been replaced.
Next to privacy measures different metadata have been collected. As a minimum we envisaged to find out for each SMS a particular contributor's age, gender, place (city, region) and country of residence. Moreover, all contributors were asked to send their email address for the iPad raffle (Section 3.2) but this information was not added in the metadata to ensure privacy.
How the metadata has been gathered and how the anonymisation has been carried out differs depending on the way in which messages have been contributed to the corpus. This is explained in closer detail below.

Android application. With the application a time and date stamp from each original, sent message are automatically added to the list as well as a unique identifier replacing the original recipient's phone number. Before sending the list to the SoNaR SMS corpus contributors could still modify or remove text messages. They received the following instructions: "To protect your privacy, we are removing sensitive information in your SMS. This process is done on your device, so your SMS is not sent to our server yet. Despite this process, you may want to have a look at the messages below and remove messages you do not wish to donate. To do this, just remove the text between the dividing lines (----)." In this draft email the contributor was asked to add gender, age and hometown. In all except two cases, the contributor indeed provided the metadata.
Further anonymisation was performed automatically by replacing sensitive data, including dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, street numbers, etc.), e-mail addresses, URLs, and IP addresses. All sensitive information is replaced with corresponding semantic placeholder codes, as shown in Table 1. Any detected e-mail address, for example, was automatically replaced by the code (EMAIL).

Original	**Example**	**Code**
E-mail	name@gmail.com	(EMAIL)
URL	www.google.com	(URL)
IP address	127.0.0.1	(IP)
Time	12:30	(TIME)
Date	19/01/2011	(DATE)
Decimal	21.3	(DECIMAL)
Integer (> 7 digits)	40000000	(#)
Hyphen-Delimited	12-4234-212	(#)
Alphanumeric	U2003322X	U(#)X

Table 1: Anonymisation Replacement Codes

SMS back-up file. This file is generated by dedicated software belonging to the mobile device and sent to the SoNaR SMS corpus by the contributor. On the website instruction were provided for generating a back-up file with a Nokia or iPhone. Contributors were free to remove SMS messages before uploading the file to the SoNaR mailbox or online dropbox. They were given the instruction not to modify the messages, but we cannot guarantee that contributors did not change the content of the messages. No automatic anonymisation was done for the export files.
Depending on the nature of the original mobile device, there is information available about the time and date stamp and the identity of the original recipient. Metadata was added by the contributor when uploading messages to the SoNaR dropbox. In case of an email contribution, metadata details were asked for afterwards. In all cases, metadata has been gathered in this way.

Online submission form. Here, SMS messages had to be manually copied by the contributor. On the SoNaR website, an online submission form was made available. Contributors were asked to copy six SMS messages from their SMS outbox (containing only 'sent' SMS messages) and to fill in their gender, age, country and town of residence. After submitting the form, a box was shown saying 'click here if you want to add more messages'. By clicking the box, the contributor was sent back to the submission where he/she could add more messages.
No automatic anonymisation was carried out for these messages, selecting appropriate text messages was left to the contributor's proper judgment. As a consequence no

[8] Due to stricter security rules in the design of Apple software, building a similar App for iPhones proved to be much more complicated. Exporting the SMS messages from an iPhone is only possible after connecting it to a computer.

time and date stamp are available for the original text messages, nor information about the original recipient's identity.

3.2 Promotion

To keep promotion costs within the project's budget limits, promotion campaigns were characterized by seeking so-called *free publicity*. Two campaigns were launched, one in The Netherlands and one in Flanders (Dutch-speaking part of Belgium).
Press releases were sent to various local and national newspapers, radio and television stations and (scientific) journalists, with the help of different university's communication offices. As a result, both in The Netherlands and Flanders small announcements were placed in several national daily newspapers. Interviews on local and national radio stations and one TV news item on a local television station were also devoted to the SoNaR SMS project.
Various researchers, active in the field of science or linguistics drew attention to the project through social media channels by adding a link to the SoNaR Facebook or Twitter page. Moreover, flyers were distributed among students on two university campuses during peak hours and professors were asked permission to give a five-minute pitch during classes at the Faculty of Arts of two Dutch universities. Fellow researchers were encouraged to bring the campaign to the attention of people in their environments (cf. the 'snowballing technique' described in Sanders & van den Heuvel, 2001).
We envisaged to reach a broad audience but because most efforts were located around our proper interest fields there might be some bias towards more educated people. This is further investigated in Section 5.2
Because we required a relatively large effort from people to actually contribute SMS messages, it was decided to put two Apple iPads up for raffle among all contributors (one in Flanders and one in the Netherlands).

3.3 Processing

After collection, text messages have been processed in order to incorporate them in the new media subcorpus of the SoNaR corpus. Data received by the Android app or the online submission form were assembled in one file. The SMS back-up export files, however, consisted of various formats and contained different character encodings, which complicated further processing. Because of this it was decided to only include files with more than 200 messages in the corpus.
SMS messages from a single contributor have been placed in the same file. All data has been converted to the FoLiA XML format[9] and tokenized with UCTO[10]. The tokeniser was adapted for social media in such a way that it recognizes e.g. emoticons. In total 52,913 messages have been collected amounting to 723,876 tokens (this amount should be placed into perspective because of the nature of this language variant).

[9] http://ilk.uvt.nl/folia/ .
[10] http://ilk.uvt.nl/ucto

4. Promotion vs. Collection

The Dutch campaign started on September 14 and the Flemish one on September 29, 2011. Both campaigns were finished on December 1, 2011. During these 12 weeks 52,913 SMS messages have been submitted by 272 contributors (147 Dutch, 125 Flemish).
Dutch donators contributed 31,586 text messages (i.e. on average 215 SMS messages per contributor) and Flemish donators 21,32 (on average 171 SMS messages per contributor). The lower average number of Flemish contributors can be explained by a lower number of contributors using the Android app (this is further discussed in Section 5.3).
The effect of various promotion activities can be roughly measured by counting the number of new contributors submitting SMS messages each day. The following effects were observed:

- For every 100 flyers, distributed on the campus, one new contributor was persuaded to donate text;
- Short presentations to groups of students resulted in a reaction from approximately one out of 40 attending students;
- Bulk mailing to students from the participating universities resulted in approximately 10 contributions for every 1,000 receiving students;
- The effect from articles in national newspapers, several radio stations mentioning the project and attention on Twitter, Facebook and other websites – all this concentrated in two or three days of publicity – both in Flanders and The Netherlands resulted in approximately 10 reactions. The slower, indirect effect such as familiarity with the project, generated by this publicity however, is not measured.

Considering the balance between the efforts put into the campaigns and the actual results, some conclusions can be drawn. Overall, we observed that free publicity, created by sending a press release in the network of the coordinating institutions, caused a strong and widespread effect. Short explanations for groups of students seem better than distributing flyers, probably because attendees estimate the contribution more trustworthy. The effect of flyers, though, is possibly more indirect which makes it hard to measure.
It can thus be highly recommended to use the expertise of the communication offices of research institutes and social media, such as Facebook, Twitter and LinkedIn. Moreover, because these media are more intensively used by smartphone owners and because of the availability of the Android app, these people are more likely to contribute larger amounts of SMS messages to the corpus (Belleghem, Eenhuizen & Veris, 2011).

5. Corpus Characteristics

In this section some characteristics of the SoNaR SMS corpus are described based on statistics. We focus on the population's distribution, the demographic properties of the contributors and the number of messages submitted versus the submission channel.

5.1 Messages per contributor

From descriptions of other SMS corpora, it is known that usually a small number of contributors contributed the bulk of the messages (Chen & Kan, forthcoming). This appears to be the case in the SoNaR corpus as well. More specifically, in our corpus 62.5% of the contributors submitted fewer than 10 messages, while the average number of SMS per contributor is 194.

The cause of this skew can be related to the different character of each collection method. As described in Section 3, the online submission form was more suitable for small amounts of SMS to be uploaded. Due to the form's design, people were likely to upload a multiplication of six messages.

With the Android app, however, both small and large numbers of messages could be contributed with a little effort. Though creating an SMS back-up file requires more effort from the contributor, this type of contribution may be more likely among frequent mobile phone users, who in turn contribute a larger amount of messages.

5.2 Demographic distribution

The total number of contributors divided among various age categories is represented in Table 2. We clearly see that the age categories 10-19 and 20-29 are most represented in the corpus. People in their twenties comprise 45% of the contributors and together with the 10-19 group they account for more than 70% of all contributors.

Age	# people	%
10-19	72	26.5
20-29	123	45.2
30-39	25	9.2
40-49	24	8.9
50-59	10	3.7
60-69	0	0.0
70-79	1	0.4
N/A	23	8.5
Total	272	

Table 2: Age distribution among the contributors

The reason for this high representation of young people is probably due to two factors: first, using text messages for communication is more common among young people (see Section 2), and second, the promotion of the SMS collection for SoNaR was largely done through the network of the university, which may have accounted for a relatively high number of students contributing their data.

Having a closer look at the gender distribution, which is illustrated in Figure 1, we see that the total number of women contributing to the corpus is higher than the number of male contributors. In Flanders and The Netherlands together, 174 women (63%) and 94 (34%) men contributed to the corpus (for six contributors gender metadata is missing). Noticeable is that the average number of messages per contribution for men is higher than for women (402 vs. 81). The total number of SMS contributed by the male contributors, 37,405 messages, covers 71% of the corpus.

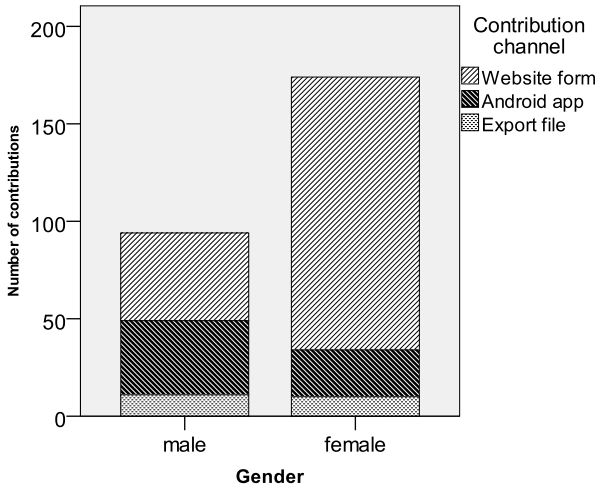

Figure 1: Number of contributions through each channel sorted by gender[11]

Another salient detail that becomes visible in Figure 1, is the larger number of contributions through the Android app by men and the relatively larger number of contributions through a website form by women. Possibly, women feel more responsible to contribute to the building of a corpus, or are more eager to win an iPad2. On the other hand, men were possibly more likely to have a smartphone in Flanders and The Netherlands in 2011. Dutch surveys confirm this: in 2010 27% of the men using the Web were consulting it through a mobile connection, against 15% of the women (Source: CBS, StatLine).

5.3 Method of contribution versus number of SMS

In Figure 2 the distribution of the number of contributions is given by contribution channel (i.e. online submission form, Android app or SMS back-up file).

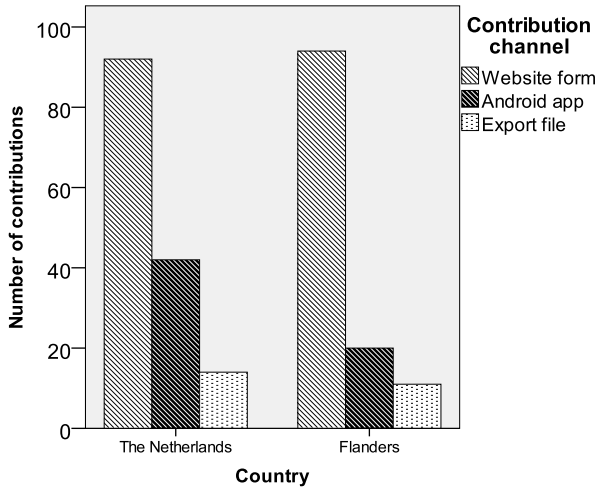

Figure 2: Distribution depending on the contribution channel in Flanders and The Netherlands.

The numbers of contributions through the website and export files are comparable between Flanders and The

[11] Two contributors with unknown gender were not included.

Netherlands, but the number of contributions with the Android app is two times higher in The Netherlands. Presumably, this is because smartphones are less common in Belgium. An explanation for this might be that Belgian telecom providers, as opposed to Dutch providers, do not provide free smartphones with phone contracts. It is likely that this makes smartphones and thus the Android platform, less popular in Flanders.

6. Conclusions and Future Work

In this paper we discussed the efforts that were invested in collecting Dutch SMS messages, completely cleared from copyrights, in the framework of the SoNaR project.

The choice of our SMS data collection method was based on other comparable projects, most notably the NUS School of Computing (Chen & Kan, 2012). Thanks to their well-documented methods and the translation of their Android app to Dutch, we were able to collect more than 52,000 SMS messages over a time span of less than three months and within a small budget.

All SMS messages have been provided with metadata information and will be distributed as part of the new media corpus within SoNaR.

Based on our findings, we advise future corpus builders to make sure that contributors are well-informed about the anonymisation of their data. The different methods described in this paper do not anonymize the data before uploading them to the corpus builders. This is also a sensible subject for legal reasons, it is very important to inform contributors about the aim of the data collection, their responsibilities in privacy issues as well as the project's responsibilities and the possibility to reject their data from the corpus in the future.

Lately, many SMS-like alternatives, such as Blackberry's Ping and the WhatsApp program for Android and Apple, seem to reduce the popularity of sending SMS messages. For that reason, it is doubtful for how long SMS will be a common way of communication.

However, this does not override the importance and relevance of an SMS corpus. There are different reasons why SMS will still be popular for many years, among others because many companies and governmental organizations have based their services and marketing on SMS.

Moreover, the challenges of processing user generated content in current NLP research and future linguistic, sociologic and technical research will benefit from corpora such as the one described throughout this paper.

Acknowledgements

This research was funded by the STEVIN programme under grant number STE07014. The App for Google Android was translated and modified for Dutch in collaboration with Tao Chen, PhD student at NUS. We would like to thank everyone who contributed SMS messages to our corpus.

References

Beaufort, R., Roekhaut, S., Cougnon, L-A. & Fairon, C., (2010). *A hybrid rule/model-based finite-state framework for normalizing SMS messages*. Proceedings of ACL 2010, Uppsala, Sweden.

Belleghem, S. van, Eenhuizen, M. & Veris, E., (2011). Social media around the World 2011. Available at www.theconversationmanager.com

Bieswanger, M., (2006). *2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different shortening strategies in English and German text messages*. Proceedings of Salsa 2006, Austin, Texas.

Chen, T. & Kan, M.-Y., (forthcoming). *Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus*. In Language Resources and Evaluation journal.

De Clercq, O. & Montero Perez, M., (2010). *Data collection and IPR in multilingual parallel corpora: Dutch parallel corpus*. Proceedings of LREC 2010, Valletta, Malta.

Grinter, R. E. and M. Eldridge. (2001): *Y do tngrs luv 2 txt msg?*, in W. Prinz et al. (Eds): *Proceedings of the Seventh European Conference on Computer Supported Cooperative Work* (pp. 219-238). ECSCW '01, Bonn, Germany. Dordrecht, Netherlands: Kluwer

Academic Publishers, pp. 219-238

Herring, S.C. & Zelenkauskaite, A., (2009). Symbolic Capital in a Virtual Heterosexual Market : Abbreviation and insertion in Italian iTV SMS. *Written communication*, 26(1), 5-31.

How, Y., (2004). Analysis of SMS efficiency. *Undergraduate thesis*, National University of Singapore.

Kasesniemi, E.-L. & Rautiainen, P., (2002*). Mobile culture of children and teenagers in Finland*.In James E. Katz & Mark A. Aakhus (Eds), *Perpetual contact: Mobile communication, private talk, public performance* (pp. 170-192). Cambridge: Cambridge University Press

Oostdijk, N., Reynaert, M. Monachesi, P., Noord, G. van, Ordelman, R. Schuurman, I. & Vandeghinste, V., (2008). *From D-Coi to SoNaR: A reference corpus for Dutch*. Proceedings of LREC 2008, Marrakech, Morocco.

Oostdijk, N., Reynaert, M., Hoste, V. & Schuurman, I., (forthcoming). *The construction of a 500-million-word reference corpus of contemporary written Dutch.* In *Essential Speech and Language Technology for Dutch: resources, tools and applications*. Springer, Verlag.

Reynaert, M., Oostdijk, N., De Clercq, O., Heuvel, H. van den & Jong, F. de, (2010). *Balancing SoNaR: IPR versus Processing Issues in a 500-million-word Written Dutch Reference Corpus*. In Proceedings of LREC 2010, Valletta, Malta.

Sanders, E. & Heuvel, H. van den, (2001). *Speaker Recruitment for Speech Databases*. In Proceedings PRASA 2001, Franschhoek, South Africa.

Sanders, E., (2012). *Collecting and Analysing Chats and Tweets in SoNaR.* In Proceedings LREC 2012, Istanbul, Turkey.

Tagg, C., (2009). *A corpus linguistics study of SMS text messaging.* Ph.D. thesis, University of Birmingham.