The Netlog Corpus
A Resource for the Study of Flemish Dutch Internet Language

Mike Kestemont, Claudia Peersman, Benny De Decker, Guy De Pauw, Kim Luyckx, Roser Morante, Frederik Vaassen, Janneke van de Loo and Walter Daelemans

CLiPS – Computational Linguistics Group
University of Antwerp
Prinsstraat 13, B-2000 Antwerp, Belgium
E-mail: <firstname.lastname>@ua.ac.be

Abstract

Although in recent years numerous forms of Internet communication – such as e-mail, blogs, chat rooms and social network environments – have emerged, balanced corpora of Internet speech with trustworthy meta-information (e.g. age and gender) or linguistic annotations are still limited. In this paper we present a large corpus of Flemish Dutch chat posts that were collected from the Belgian online social network Netlog. For all of these posts we also acquired the users' profile information, making this corpus a unique resource for computational and sociolinguistic research. However, for analyzing such a corpus on a large scale, NLP tools are required for e.g. automatic POS tagging or lemmatization. Because many NLP tools fail to correctly analyze the surface forms of chat language usage, we propose to normalize this 'anomalous' input into a format suitable for existing NLP solutions for standard Dutch. Additionally, we have annotated a substantial part of the corpus (i.e. the *Chatty* subset) to provide a gold standard for the evaluation of future approaches to automatic (Flemish) chat language normalization.

Keywords: chat language, normalization, Flemish Dutch

1. Introduction

Recent decades have brought a rapid succession of new communication technologies, including text messages or the numerous forms of Internet communication (e.g. e-mail, blogs, social media). When compared to earlier (e.g. handwritten) forms of communication, these computer-mediated technologies stand out because of an increased level of 'immediacy'. People tend to communicate more and faster, so that their writings become increasingly casual and reminiscent of oral communication. An obvious effect of these recent developments has been the wild proliferation of language variation in written communication, especially affecting surface phenomena such as spelling. Speakers generally consider their standard language inadequate for these new settings and adopt a 'glocal' language variety, displaying both characteristics from a global 'Internet language' as well as their local dialect (Androutsopoulos, 2010).

The fast-paced developments in computer-mediated communication are a main object of study in sociolinguistics (Vandekerckhove & Nobels, 2010), but balanced corpora of Internet speech with trustworthy meta-information (e.g. age and gender) or linguistic annotations (e.g. ellipsis, interjections, regional and foreign language usage, …) are scarce. The few resources that do exist (e.g. Forsyth and Martell, 2007) focus on *linguae francae*, cold-shouldering the re-emergence of vernacular features in written communication. Moreover, corpora that contain meta-data and/or linguistic annotations can also be used in computational linguistic studies such as opinion and sentiment mining. However, most software architectures for Natural Language Processing (NLP) take on a modular workflow in which low-level results from surface analyses (e.g. tokenization) are used as the input for more advanced procedures (e.g. semantic role labeling) and were not designed to cope with the intense surface variation in Internet speech. Consequently, they fall short in the early stage of the analysis (cf. Liu et al., 2010).

In this paper we present the Flemish Dutch Netlog Corpus, a comprehensive collection of 1.5 million posts, provided by the Belgian social networking platform Netlog[1] in the context of a project aimed at the detection of pedophiles in chat rooms through text analysis (Daphne project[2]). Each post contains meta-information on the user's profile, making this corpus a unique resource for computational and sociolinguistic research.

2. The Netlog Corpus

Netlog is a Belgian online social networking platform, targeting European adolescents, with over 67 million members, utilizing over 37 different languages. Members can create a profile page containing blogs, pictures, videos, events, playlists, etc. that can be shared with other members. A collection of 1.5 million Flemish Dutch Netlog posts has been obtained, amounting to ca. 19 million running tokens (i.e. words, emoticons, and punctuation marks). The posts are typically short with an average length of twelve tokens per post. For each post, we have access to meta-information about the profile of the user who posted it, such as age, gender and location (Peersman, Daelemans & Van Vaerenbergh, 2011).

The bulk of the Netlog Corpus is written in colloquial Flemish, a conglomerate of Dutch dialects spoken in the North of Belgium (Flanders). It differs significantly from Netherlandic Dutch and displays a lot more dialectal features. Colloquial Flemish currently attracts much research interest, because it adopts characteristics from local dialects as well as from Flemish Standard Dutch. It is therefore considered an 'intermediate' variety of Dutch

[1] www.netlog.com (Accessed: 14 March 2012).
[2] http://www.clips.ua.ac.be/projects/daphne (Accessed: 14 March 2012).

(Vandekerckhove & Nobels, 2010). The casual language documented in the Netlog Corpus is well suited for research into this language variety, especially because of the geographical information included in the user profiles.

This information also enables the study of posts according to their geographical distribution into the 'regiolects' of Flanders' provinces. The colloquial nature of Internet language causes the large-scale introduction of non-standard features in written language. However, in the absence of spelling norms for them, such vernacular features can take on multiple surface forms (e.g. the exclusively West-Flemish interjection in 'wi', 'whe', 'weh', …). Table 1 provides samples of Flemish 'regiolect' usage from the corpus alongside their equivalents in Standard Dutch and English.

Regiolect	Netlog	Standard Dutch	English
West-Flemish	zitn kik omeki zo verre?	Zit ik ineens zo ver?	Am I that far suddenly?
East-Flemish	est gedon	Is het gedaan?	Is it over?
Antwerp	wa hedde gij die schoene gekocht	Waar heb jij die schoenen gekocht?	Where did you buy those shoes?
Flemish-Brabant	we hebbe toch gelache zemen	We hebben toch gelachen, hoor.	We had a good laugh, didn't we?
Limburg	hou gans veel van u	Ik hou heel veel van jou.	I love you very much.

Table 1: Samples of non-standard varieties in the Netlog Corpus (Peersman, Daelemans & Van Vaerenbergh, 2011).

The Netlog Corpus displays many of the typical Internet language characteristics that have also been reported for other languages in the international literature (Baron 2003, Crystal 2001). These include the frequent omission of words, punctuation, whitespaces, and characters (cf. abbreviations and acronyms), as well as intensive spelling variation – including, but not limited to plain errors.
Moreover, expressive discourse markers such as character flooding ('hiii') or the use of uppercase characters are frequently attested. Table 2 shows examples of such phenomena in the Netlog Corpus. Flemish Dutch chat language, however, shows a number of specific features as well, such as the non-standard concatenation of tokens – especially affecting grammatical morphemes – and the associated assimilation processes. The singular first-person pronoun 'ik', for instance, is frequently attached as a clitic to verbs: 'kvind' for 'ik vind' (*I find*), 'kweet' for 'ik weet' (*I know*). Although the 'k' behaves like a clitic in these contexts, it can only be mapped to an independent lexeme in Standard Dutch.

Such processes naturally cause large-scale ambiguities in the corpus, especially for computational analyses. When applying a standard part-of-speech tagger, 'kweet' would be tagged as the simple past tense of the low-frequent verb 'kwijten' (*to acquit*), while in chat language it represents the highly frequent though non-standard form of 'ik weet' (*I know*). Consequently, it should be tagged as a combination of the personal pronoun and the present tense of the verb 'weten' (*to know*).

Variation type	Netlog example	Standard Dutch	English
Omission of characters or words	kbda nimr	Ik heb dat niet meer.	I don't have that anymore.
Acronyms	Hjg	hou je goed	take care
Character flooding	keiii mooiii	heel mooi	very beautiful
Concatenation	IkKanOokNiii ZonderU!	Ik kan ook niet zonder jou!	I can't live without you either!

Table 2: Samples of non-standard varieties in the Netlog Corpus.

3. The Chatty Subset

Because standard NLP tools fail to correctly analyze the surface forms of Internet chat language, the *Chatty* project proposes to normalize this 'anomalous' input into a format suitable for existing NLP solutions for standard language. Therefore, we have annotated a part of the corpus – the *Chatty* subset – in the context of a research project that deals with the automatic normalization of (Flemish) Dutch chat language. This *Chatty* subset will be used as a gold standard data set for the evaluation of our future approaches to normalization.

Chatty focuses on teenagers' language usage, since youngsters' Internet language has attracted most of the attention in the scientific literature (Vandekerckhove & Nobels 2010). Following the approach in Xia et al. (2005), we initially annotated a representative selection of 1,000 posts manually. In a second stage we aim to annotate the rest of the corpus automatically.

We have selected posts by authors who claim to be younger than 26 years of age. This selection has been balanced for the Flemish 'regiolects' (§2), ensuring an equal distribution over the five provinces. Only posts that contained a minimum of 15 surface tokens (words, smileys, and punctuation marks) were selected. As such, the gold standard will contain at least 15,000 annotated tokens, which is an acceptable working base for the evaluation of our work. Given that the Netlog data can also contain other text genres than chat (e.g. English-language commercials, poetry, …), we decided to flag these posts as 'non-chat' and exclude them from our annotation task. Also, Netlog posts can include previous entries in a conversation between several users. Since we do not have the meta-information of the other users, we only use the last entry of each post.

The normalization task we envisage serves a pragmatic goal: the output of the system will be considered appropriate, if it can be parsed correctly by existing software (e.g. *Frog*: van den Bosch et al. 2007). The normalization of spelling variation or the restoration of unbound clitic morphemes (cf. 'kweet' > 'ik weet'), for

instance, are important aspects of this normalization task. Our normalization task, however, does not involve translating regional lexical variants – word stems that are not part of Standard Dutch – into a canonical equivalent[3]. Our normalization is restricted to 'rewriting' surface forms (cf. transliteration), involving the normalization of orthography and morphology, but not of lexis or syntax. Each post was annotated independently by at least two annotators. Furthermore, any annotator disagreements that arose, were resolved during a final adjudication phase.

The proposed annotation style is simple, yet effective and uses Excel-files (illustrated in Table 3). A first data column ('Anomalous') contains the original tokens as present in the data, with posts split along whitespaces. A second column ('Tokenized') holds the automatically tokenized version of the anomalous column, so that the annotators only had to correct the tokenizer we built for this task, instead of providing the tokenization manually. This way, each row contains a single raw token instance. The third column ('Normalized') contains the normalized forms, potentially mapping a single complex token to multiple normalized tokens and vice versa.

We then introduced six additional 'flag' columns that provide extra annotation labels to the *Chatty* subset. These are listed in Table 4: the 'INTJ' (for meaning-poor discourse markers such as conversational initialisms, smileys and other interjections that are not informative enough to be parsed), 'Regional' (indicating the presence of regional or dialectal lexical morphemes), 'Foreign' (for loan words), 'NE' (for named entities), 'Ellipsis' and 'Non-chat' (for other text genres such as poetry or advertisements) columns.

Next, guidelines were developed that describe the details of the annotation. These entailed correcting spelling errors and spelling variations, including capitalization in the case of named entities, and correcting the non-standard use or absence of whitespaces (e.g. in concatenated forms (§2, Table 2)). If two or more consecutive tokens needed to be merged (e.g. in row *n*: 'e' and row *n +1*: 'mail'), the annotators appended a plus-sign ("+") immediately (without whitespace) before and/or after the tokens that needed to be merged in the Normalized-column. Missing punctuation marks were not inserted.

Although we did not change the underlying core of morphemes (e.g. word stems), we did indicate the use of regional stems and normalized the spelling and inflection of these words to their standard form. For example, chatters often reduce the standard infinitive suffix 'en' to 'e' or 'n' depending on their own dialect (Vandekerckhove and Nobels, 2010) and use e.g. 'zegge' or 'zeggn' instead of the standard 'zeggen' (*to say*). However, the typically Flemish third person singular pronouns 'ge', 'gij' and Flemish singular use of 'u' and 'uw' are grammatically correct and will therefore not be adjusted to the standard Dutch forms 'je', 'jij', 'jou' and 'jouw', but only flagged as 'Regional'. The spelling variants 'gy' and 'jy' on the other hand, were corrected to 'gij' and 'jij' (*you*).

Additionally, all abbreviations were expanded, also when they reflected multi-word units (e.g. 'idd' to 'inderdaad' (*indeed*) and 'hjg' to 'heb je graag' (*like you*)). Also, character flooding (§2) was corrected, including when it affected INTJs (e.g. 'looool' to 'lol'). Foreign tokens (including multi-word units) were marked as 'Foreign' and were only adjusted for flooding, capitalization and abbreviation correction.

Furthermore, we used indices to indicate which tokens on a particular line should be 'flagged' in these columns. We also incorporated the possibility to include ranges. For example, when confronted with a concatenated form that is normalized to three regional forms followed by one standard word, the annotators could indicate this by adding '1-3' or '1,2,3' in the 'Regional' column. The example on the next page in Table 3 (which for reasons of clarity includes an English translation) illustrates the positive effects of the normalization on the output of existing state-of-the art parsing software for Dutch.

4. Conclusion

Internet language presents an important challenge for present-day (computational) linguistics. Apart from its (theoretical) relevance in e.g. sociolinguistics, it currently impedes a number of industrial applications, such as opinion and sentiment mining, on a more practical level. Both the Netlog Corpus and its annotated *Chatty* subset will be valuable resources for such future research. Chatty can serve as a gold standard for the evaluation of future approaches to automatic (Flemish) chat language normalization which will be developed in the CLiPS group. Note, however, that the (relatively small) *Chatty* subset is not intended to be used as training material for supervised normalization strategies. In a world where adolescent language changes in a quick-paced fashion, we believe that machine learning approaches for the normalization of such language are needed which require minimal supervision, in order to reduce the cost and effort of manual annotation.

5. Acknowledgements

This study has been carried out in the framework of the DAPHNE project and the BIOGRAPH project, both funded by the University of Antwerp in Belgium; the deLearyous and the ALADIN project, both funded by the Belgian government agency for Innovation by Science and Technology (IWT) and the Language Geography project, funded by the Research Foundation – Flanders (FWO). Guy De Pauw and Mike Kestemont are funded as fellows of the Research Foundation – Flanders (FWO). The research of Luyckx and Daelemans is partially funded through the IWT project AMiCA: Automatic Monitoring for Cyberspace Applications. We also thank Netlog and Mollom for supplying the data needed for constructing this corpus and the reviewers for their comments that helped improve the manuscript.

[3] Although they were annotated as such for future research purposes.

#	Anom.	Tok.	Norm.	English	NE	INTJ	Ellipsis	Regional	Foreign	Non-chat	Anom. POS	Anom. DP	Norm. POS	Norm. DP
1	hoop	hoop	hoop	hope			1				Noun	subj 2	Verb	ROOT
2	egt	egt	echt	really							Verb	ROOT	Adj	mod 1
3	dak	dak	dat	that							Noun	dir.obj 2	Conj	verb.compl 1
4			ik	I									Pers.Pron	subj 7
5	niiii	niiii	niet	not							Noun	subj 7	Adv	mod 6
6	ziek	ziek	ziek	ill							Adj	predc 7	Adj	predc 7
7	zen	zen	ben	am				1			Verb	mod 2	Verb	body 3
8	:p	:p				1					Punct	punct 7		

Table 3: Example of an annotated *Chatty* post with *Frog* POS tags and dependency parses (DP) of the anomalous and normalized version.

6. References

Androutsopoulos, J. (2010). Localizing the Global on the Participatory Web. In N. Coupland (Ed.), *The handbook of language and globalization*. Oxford : Wiley-Blackwell, pp. 203--231.

Baron, N.S. (2003). Why email looks like speech : proofreading, pedagogy and public face. In J. Aitchison, & D.M. Lewis (Eds.), New Media Language. London, New York : Routledge, pp. 102--113.

Crystal, D. (2001). *Language and the Internet*. Cambridge, NY, USA: Cambridge University Press.

Forsyth, E. and Martell, C. (2007). Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the 2007 IEEE International Conference on Semantic Computing*, Washington, DC, USA: IEEE Computer Society, pp. 19--26.

Liu, X., Li K., Han B., Zhou M., Jiang L., Xiong Z. and Huang C. (2010). Semantic role labeling for news tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 698--706.

Peersman, C., Daelemans, W. and Van Vaerenbergh, L. (2011). Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. New York, NY, USA: ACM.

Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. F. van Eynde et al. (eds.), In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*. Utrecht, The Netherlands: LOT, pp. 99--114.

Vandekerckhove, R., and Nobels, J. (2010). Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers. *Journal of Sociolinguistics*, 14(5), pp. 657--677.

Xia, Y., Wong, K. and Li, W. (2005) Constructing a Chinese chat language corpus with a two-stage incremental annotation approach. In *Proceedings of NTCIR-5 Workshop Meeting.* Tokyo, Japan.