

Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus

Martin Reynaert¹, Nelleke Oostdijk², Orphée De Clercq³,
Henk van den Heuvel², Franciska de Jong⁴

ILK, Tilburg University¹; CLST, Radboud University Nijmegen²;
University College Ghent/Ghent University³; HMI, University of Twente⁴

Abstract

In The Low Countries, a major reference corpus for written Dutch is currently being built. In this paper, we discuss the interplay between data acquisition and data processing during the creation of the SoNaR Corpus. Based on recent developments in traditional corpus compiling and new web harvesting approaches, SoNaR is designed to contain 500 million words, balanced over 36 text types including both traditional and new media texts. Beside its balanced design, every text sample included in SoNaR will have its IPR issues settled to the largest extent possible. This data collection task presents many challenges because every decision taken on the level of text acquisition has ramifications for the level of processing and the general usability of the corpus later on. As far as the traditional text types are concerned, each text brings its own processing requirements and issues. For new media texts - SMS, chat - the problem is even more complex, issues such as anonymity, recognizability and citation right, all present problems that have to be tackled one way or another. The solutions may actually lead to the creation of two corpora: a gigaword SoNaR, IPR-cleared for research purposes, and the smaller - of commissioned size - more privacy compliant SoNaR, IPR-cleared for commercial purposes as well.

1. Introduction

Within the STEVIN project SoNaR work is underway directed at the construction of a major Dutch reference corpus. The SoNaR project started on January 1, 2009 and runs until December 2011. Its aim is to develop a 500-million-word balanced reference corpus for contemporary (1954-present) written Dutch. The corpus will comprise data originating from Flanders, the Dutch speaking language area in Belgium, and the Netherlands. The texts will be balanced according to the approximate numbers of speakers of Dutch in both countries, i.e. one-third of the texts should have originated in Flanders, two-thirds in the Netherlands. Beside texts from the more conventional text types, data from new media such as chat, SMS, internet fora and email will be included.

The SoNaR Corpus is an exponent of a new generation of corpora that is emerging. As storage capacity is no longer in the way of constructing ever larger corpora, and there is an abundance of texts associated with the new media that begs investigating, it is clear that we need to re-address issues of corpus design and compilation.

In this paper we describe how in the present landscape and taking advantage of various advantageous conditions, we are implementing a corpus design that is unprecedented. We also want to share our findings regarding the acquisition of texts and the settling of Intellectual Property Rights (IPR). We will show how some decisions taken with respect to the acquisition of the texts have serious ramifications for the level and means of processing later on, as well as on the general usability of the corpus and on the targeted user base.

The structure of the paper is as follows. First, in Section 2, the SoNaR Corpus is placed within recent developments in corpus design and compilation. Next, in Section 3, we briefly compare our approach with traditional corpora and also with more recent corpora built from the web. Sections 4, 5 and 6 deal with the interplay between data collection,

IPR and processing. In Section 7 we take a brief look at corpus quality control and validation. Section 8 concludes this paper.

2. The SoNaR Corpus

A corpus is usually compiled with a specific goal in mind. The stated goal of the SoNaR Corpus is to fill a major gap in the tool kit of linguists and researchers in the field of language technology, viz. a widely available, large and balanced reference corpus of written Dutch. Thus the corpus should be a well-structured, balanced collection of text samples tailored to the different uses to which the corpus is going to be put. The contents of the corpus as well as the nature of the annotations to be provided are largely determined by the needs of ongoing and projected research and development in the fields of corpus-based natural language processing. Applications such as information extraction, question-answering, document classification, and automatic abstracting will benefit from the large-scale analysis of particular features in the corpus. Apart from supporting corpus-based modeling, the corpus will constitute a test bed for evaluating applications, whether or not these applications are corpus-based. As a consequence, in terms of the design of the corpus, the SoNaR Corpus intentionally deviates from previous corpora and data collection initiatives.

2.1. The SoNaR Corpus in the light of the state of the art in corpus building

When we consider the developments in compiling written corpora through the years, these can be summarized as follows.

- *Corpus size* has grown from one million words (MW) in early corpora such as the Brown and LOB (Lancaster-Oslo-Bergen) corpora in the 1960s to 100 MW for the British National Corpus (BNC)¹ in the late

¹Cf. <http://www.natcorp.ox.ac.uk/>

1990s to 400+ MW (Bank of English² and COCA, the Corpus of Contemporary American English³) in the last decade. With SoNaR we aim for a corpus of minimally 500 MW.

- A change in the composition of corpora as regards *text length*: traditional corpora are typically collections of samples of written and spoken language, e.g. the early Brown and LOB corpora, which comprise 500 samples of 2,000 words each. Through the years sample size increased but there remained a tendency to collect samples (cf. the BNC) rather than full texts. The SoNaR Corpus will include full texts, ranging in size from the character limit imposed by SMS to reports covering thousands of pages.
- A change in the composition of corpora as regards *content*: traditional corpora include publications in traditional media (books, magazines, newspapers, journals, leaflets, reports). Texts available through the internet and new media only started to appear in corpora as from the mid-1990s. The SoNaR Corpus will include a large variety of genres and media, both traditional and novel.
- Traditional corpora were perceived as *static collections*, which is related to the idea of a design underlying a balanced composition of the corpus. More recently there is a tendency to view corpora as *dynamic*. Thus the Bank of English contains 524 million words and continues to grow, while the Corpus of Contemporary American English is already 400+ MW and is updated every six to nine months. The SoNaR Corpus can potentially become dynamic as various agreements with text providers foresee ongoing text delivery.
- *IPR* has remained problematic and has affected the availability of corpora (cf. the Bank of English and the Corpus of Contemporary American English as well as the WaCky corpora⁴) and the balancedness of corpora. In the SoNaR project we aim to settle IPR for all texts included in the corpus to the fullest extent possible.

2.2. (Written) Dutch language corpora

For Dutch a number of corpora are already available.⁵ For contemporary written Dutch these include

- the Eindhoven Corpus; a collection of Dutch texts that were published between 1960 and 1976. The corpus comprises some 768,000 tokens.
- several corpora compiled by the Institute for Dutch Lexicology (INL), e.g. the 5 MW Corpus (1994) which is a collection of texts from books, magazines and broadcast news covering the period 1970-1994, the 27 MW Newspaper Corpus (1995) which includes

newspaper articles from the NRC Handelsblad published between January 1994 and April 1995, the 38 MW Corpus (1996) which consists of three main components, viz. a mixed component (including materials from 1970-1995), a newspaper component (Meppeler Courant, 1992-1995), and a legal component (1814-1989), and the PAROLE Corpus (2004), a 20 MW collection of contemporary Dutch texts (mostly from newspapers, magazines and books published between 1982 and 1998).

- the Twente Nieuws Corpus (TwNC⁶ which includes over 300 MW of newspaper texts, teletext subtitling and autocues of broadcast news shows and news texts downloaded from the WWW.

However, the corpora that are presently available do not (individually nor collectively) suffice to satisfy the need for sufficiently large and varied amounts of data that have been cleared for IPR so that they can be widely used not just for studying individual words or phrases but also for training language models and such.⁷

2.3. SoNaR Corpus design

As already observed above, the SoNaR Corpus is intended as a reference corpus that can be used by linguists from various subdisciplines as well as researchers working in the field of language technology. In designing the SoNaR Corpus we did not simply want to replicate the design of a corpus such as the BNC or COCA.

Table 1 lists the overview of text types and projected amounts of word tokens being or to be incorporated in SoNaR.

3. Implications of the SoNaR approach to corpus building

There have been important developments in corpus building from the web in recent years. The main differences between these and the traditional corpora have been well documented in a range of papers by the WaCky group (Baroni et al., 2009); (Evert, 2008); (Baroni and Kilgarriff, 2006). Our approach treads a middle ground between the traditional corpus building initiatives and the web harvesting approach. We think that a web-as-corpus approach would not deliver a corpus that would serve the many purposes pursued with the SoNaR Corpus.

Table 2 lists the most salient differences and main ramifications of our approach versus a web-based approach.

The fact that we settle IPR issues to the largest extent possible has immediate consequences for the corpus itself. We next enumerate first the main advantages we see in our approach, next the disadvantages.

1. Having settled IPR, we can make the corpus available to anyone willing to sign the licence for users which clearly delineates the user's rights and obligations.

²See <http://www.collins.co.uk/books.aspx?group=153>

³See <http://www.americancorpus.org/>

⁴<http://wacky.sslmit.unibo.it/doku.php>

⁵The corpora can be obtained from the Dutch HTL Agency: <http://www.inl.nl/nl/corpora/>

⁶See <http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html>

⁷The 300 MW TwNC does allow for training language models. Unfortunately, the data offers little text type variation.

Text types	SoNaR
Written to be read, published, electronic	
Discussion lists	2.5 MW
E-books	5 MW
E-magazines	25 MW
E-mail (spam)	2.5 MW
Newsletters	2.5 MW
Press releases	10 MW
Subtitles	10 MW
Teletext pages	50 MW
Websites	50 MW
Wikipedia	20 MW
Written to be read, published, printed	
Abstracts, summaries	10 MW
Books	75 MW
Brochures	5 MW
Newsletters	2.5 MW
Guides, manuals	5 MW
Legal texts	2.5 MW
Newspapers	50 MW
Periodicals, magazines	10 MW
Policy documents	5 MW
Proceedings	10 MW
Reports	5 MW
Surveys	2.5 MW
Theses	2.5 MW
Poems	no set target
Written to be read, unpublished, electronic	
Chats	25 MW
E-mail (non-spam)	50 MW
Minutes	10 MW
SMS	5 MW
Written assignments	10 MW
Blogs	no set target
Written to be read, unpublished, printed	
Theses	10 MW
Written to be read, unpublished, typed	
Minutes	10 MW
Written assignments	10 MW
Written to be spoken, unpublished, electronic	
Autocues	2.5 MW
Written to be spoken, unpublished, typed	
News scripts	2.5 MW
Texts for the visually impaired	2.5 MW

Table 1: SoNaR Corpus design. Listed are 34 text types for which we have a set word token target. We further include blogs and poems, for which the design specifies no specific target, bringing the total to 36 text types

2. Since we need to invest effort in settling the IPR, we invest prior effort in choosing the likely prospects for text donation. Largely through the web, we study which prospective donators hold interesting text collections. Based on both available text quantity and relative quality we initiate negotiations with the prospective donator. Our use of the word ‘donator’ here in fact implies that we never pay in order to settle IPR. We not only have no budget for this, but we also take

Advantages
Greater availability through IPR-clearance
Text type diversity (36 text types)
Controlled synchronicity
Balancedness
Metadata storage
Easy extraction of subcorpora
Control over processing
Control over text quality
Disadvantages
Money and time consuming
Ambiguous text classification
Text format diversity
Need for robust conversion tools
Various data transfer channels
Changing nature of text material
No hypertext information
Non-running text not included in XML format

Table 2: Advantages and disadvantages of the SoNaR approach versus a web-as-corpus approach

the principled stand that the SoNaR Corpus should be available for free for research purposes.

3. Once negotiations have been successfully completed, we can automatically download the text materials acquired in a fully targeted fashion. Instead of requiring special-purpose web harvesters as in the WaCky approach, we can suffice with the Unix tool ‘wget’ for targeted automatic download. A direct consequence of this is that we have control over the synchronicity of the written texts we collect. Given the fact that currently ever more previously paper-locked archives are being digitalised and put online after conversion into electronic text by means of the highly error-prone Optical Character Recognition (OCR) technology (Reynaert, 2008), we think that this constitutes a major possible pitfall as regards the lexical quality of web harvested corpora, even if there are attempts to try to deal with this problem (Ringlstetter et al., 2006).
4. The WaCky corpora available to date are all based on html format documents only, to the best of our knowledge. In contrast, given the very wide range of text types we collect, we acquire texts in all sorts of electronic formats. We in fact do not include OCRed text. For all the formats we do handle, we need to first identify the best possible conversion steps to be taken and next adapt the suite of conversion tools we have at our disposal to the particular batch of texts acquired. The latter is more often than not in part a result of the physical layout of the text and the whereabouts of the metadata within the documents, which we can then collect and treat in a far more appropriate manner than is possible with the fully automatic collection and conversion process in a web-as-corpus setting. Having expended the effort of manual adaptation of our tools to a particular batch, we can then direct our attention elsewhere

while the whole batch is converted in an appropriate fashion by our computers, however large the batch in reality. From a single internet forum, we have in this way collected over 480 million word tokens of IPR settled Flemish Dutch text, even if the SoNaR design specifies only one-third of 2.5 MW of Flemish internet forum text. This and similar huge amounts of surplus text will nevertheless be made available for those researchers whose work on e.g. language modelling requires text in bulk. The levels of linguistic annotation added to this surplus will be dependent on the availability of processing power.

5. Given that we fastidiously collect all the metadata available, the way is open for the user of SoNaR for extracting subcorpora on the basis of specific criteria. The gateway to doing this is in fact the database in which we also store the metadata (apart from inline in the xml files themselves).
6. Having settled IPR issues with a particular donator in either country, with an eye on the balancedness of the corpus, we try to identify a similar organization in the other country having similar, possibly domain specific, documents and try to obtain permission from them as well.
7. In exceptional cases we can benefit greatly from work being done in other projects. One such is the current endeavour by (Marx and Schuth, 2010) with whom we collaborate. From the DutchParl corpus we will be able to draw many times the word quota for Legal Texts and Policy Documents and further linguistically enrich them in SoNaR.

Our approach does have its disadvantages.

1. In contrast to web harvested and fully automatically processed corpora, our approach is of course expensive in terms of time and money. Given the broader use we think our corpus can be put to, this should be a good investment.
2. The conversion process does not allow for retaining the hypertext information. So possibly useful or valuable information about the interlinkedness of documents is lost. If ever we come to regret this, a possible way out may be available in the Internet Archive. If this continues to exist as it does today, it may allow the researcher in need of this information, on the basis of the metadata about our web download that we store in the metadata, to retrieve the original html files.
3. Non-running text is not preserved in our xml format. The decision has been to not retain non-running text, so we discard what is not discourse. This includes tables, headers and footers and may include footnotes and foreign language text longer than a phrase within a sentence.

4. Acquisition and IPR

Collecting 500 million words of written Dutch from at least two countries, spread over 36 text types and coming from

both traditional and new media (Oostdijk et al., 2008) constitutes a challenging data collection task.

Data collection started in the first phase of the project and will continue until the end. The envisaged text types are divided into three groups that will be released at yearly intervals during the project. A first release of 156 MW was delivered in June 2009 and we are currently working towards the second one scheduled for June 2010. SoNaR Release 1 is available for research purposes from the Dutch Human Language Technology Agency (HLT Agency)⁸.

In practice, collecting different text types entails different strategies for their acquisition: a lot of the text types bring their own specific complications. In extension of the useful 'library' metaphor introduced by (Evert, 2006), we would also like to distinguish between what could metaphorically be described as 'public' and 'private' libraries. In contrast to expectations major parts of these libraries are paperlocked even today. We are e.g. negotiating with a major book publisher. As it turns out, the bulk of their holdings are not available in digital format at all. Another problem are the bulk of the SMS sent or chats conducted daily: these are born-digital, but due to technical obstacles and prevailing copyright law, these are effectively also not part of the 'public library'.

Because the SoNaR Corpus will be made available for the entire research community, considerable efforts are being put into IPR settlement. This is done in close cooperation with the HLT Agency. To this end we use two licence agreements: a contract that allows for commercial use and one that does not. These are in fact two shorter and to potential donators more palatable versions of the licence agreements that were used during the creation of another STEVIN-funded corpus, the Dutch Parallel Corpus (De Clercq and Montero Perez, 2010). Our experience has taught, as common sense should tell, that the less juridically complicated the agreement, the faster donators are inclined to give their consent and sign. During acquisition talks, it may occur that the prospective provider cannot agree with the standard terms, after which a customised version has to be drawn up and negotiations restarted.

New media text types require an even more flexible IPR settlement procedure, viz. one that can be handled electronically and does not require sending back and forth by land mail. For the donation of SMS, email and other new media we are looking into setting up an electronic drop box. The assumption there will be that anyone donating through this drop box implicitly consents to the reuse in the corpus. Further we are investigating whether for SMS one might accept that the very shortness of the messages implies they fall under citation law. For chats and short messages on public display online one might wonder why copyright should be in force and for these messages put online under an assumed name or an alias one might very well wonder whether this in itself does not legally sufficiently ensure the poster's privacy so that further anonymization is not required.

During the course of the acquisition and IPR settlement process we have gradually built up a manual which covers

⁸The Dutch-Flemish agency for management, maintenance and distribution of Dutch digital language resources. See <http://www.inl.nl/nl/corpora/>

all aspects of prospecting, contacting the prospect, negotiation, drawing up the IPR agreement. The manual will be made available online on the SoNaR website⁹ so as to help future corpus builders. Each specific negotiation issue is illustrated with real-life correspondence, for which we provide the translations in English.

5. Processing the texts

Text types bring their own issues for the collection of the data, but also for the processing of the data. In our experience, each batch of texts brings its own processing requirements and issues.

For each text we build three xml versions.

- The basic XML version has the text in paragraph delimited format.
- The second gives a one word per line sentence-split and tokenized format.
- The third has part-of-speech tags and lemmas added.

Each of these has an IMDI¹⁰ header in which the metadata is stored inline. For each of these there is an additional file with the XML validation report.

These files are currently put in a flat directory structure, one for each text type. We also maintain an online database in which all the metadata from the headers and the summary of the XML validations is stored. This database in fact documents the corpus and facilitates corpus access and the selection of subcorpora.

6. Interplay between acquisition and processing

Our IPR licences promise the providers that, while commercial users of the corpus may use the data for developing new products, they will not be allowed to use, copy or make available the data in a recognizable form. Nevertheless, we build a corpus of flowing discourse, which by its very nature implies that the texts can be studied as texts, in plain language: that the text is readable, at least if one knows how to read it. This apparent contradiction is solved by our reliance on inline XML, in tandem with the sheer volume of the corpus.

This also raises questions about what text types one should go for and in what quantities. It is all too clear that one-to-one communications such as SMS are very hard to acquire, both in terms of IPR and in terms of handling. Conversely, there are some text types which may well display very similar language characteristics, but which are one-to-many communications and are typically available online. For SMS, for instance, there are the public Twitter 'tweets'. These can be harvested online, but require highly accurate language filtering, for which we rely on 'TextCat'¹¹.

Settling the IPR with an internet forum has proven remarkably easy provided the forum's terms of use state that the rights to the posts are transferred to the forum and because

forums are based on database systems the data are easily and accurately convertible. In one particular case, the owners of the internet forum let us have an offline copy of the contents of the archive but withheld specific database columns, among which the one holding the posters' aliases or names, for privacy reasons. Lacking this list of names, we cannot in fact further anonymize this forum. Given that one has downloaded an entire internet forum, one can in fact on the basis of the list of names or aliases use these to effect more thorough anonymization. This was the case with the 480 MW Flemish forum mentioned above. During processing, on the basis of the posters' aliases, we counted their number of posts. Using this information we then ranked them in descending order of posting, dubbing the most prolific poster 'Poster00001'. The text of the posts was then searched for the original names or aliases and anonymized by replacement with their new, ranked alias. Ensuring anonymity in this way, however, renders some subparts of the corpus utterly useless for e.g. Named Entity Recognition. This might justify our choice to include both the non-anonymized and anonymized version in the corpus and possibly even place restrictions on the non-anonymized version so as to ensure that it is only available for particular types of research.

This issue highlights yet another aspect of corpus building. SoNaR will be more than just one 'widely useful' corpus, what seems to emerge is a collection of many subcorpora - some the size of SoNaR - that may or may not be useful for the purposes of specific research.

A corpus of this kind will probably be used for many years to come. For this reason, we do not wish to impose our own, necessarily limited views on what a corpus should have to offer. We nevertheless have to keep an eye on quality, which is why we take a brief peek at our quality control methods and evaluation plans in the next, prefinal, section.

7. Quality control and evaluation

Quality control is an essential element in the production of any language resource. It should take place all along the production time line of the resource, rather than being put as a final check at the very end of corpus completion. By allowing quality assessments at the beginning of the production process mistakes can be repaired at an early stage that would be disastrous when discovered at the end of the production process. For that reason, we consider the following quality control mechanisms as essential for the SoNaR project:

1. A prevalidation phase in which corpus design and integrity of data annotations is checked from an early stage throughout production.
2. Safeguarding the quality of the end product
3. Monitoring the external validation

7.1. Prevalidation

For the prevalidation task there will be a series of quality assessments of data annotation from an early stage of the project. The prevalidation is directed towards the correctness of the annotations and comprises:

⁹<http://lands.let.ru.nl/projects/SoNaR/home.html>

¹⁰<http://www.mpi.nl/IMDI/>

¹¹<http://www.let.rug.nl/vannoord/TextCat/>

- the orthographical annotation
- morphological annotation, lemmatization, POS tagging

More manual effort will be needed for the quality assessment of the semantic annotations (named entities, coreference, semantic roles, spatial and temporal relations). These semantic annotations are part of a separate SoNaR work package described in a companion paper (Schuurman et al., 2010).

7.2. Safeguarding the quality of the end product

Before the SoNaR Corpus can be submitted to final validation, it needs an internal quality assessment. To that end, the following parts of the corpus will be checked:

- Corpus documentation
- Corpus design and completeness
- Corpus annotation

Except for the assessment of corpus documentation and the corpus design, these checks will be carried out automatically mainly at a formal level. This permits us to go through the full corpus and check its integrity and completeness at all annotation levels. Scripts will be written to perform these checks.

7.3. Final validation

For monitoring the final validation by an external party a proper instruction is needed in order to carry out the validation of the SoNaR Corpus effectively and efficiently. To that end a document must be written that contains a clear instruction of the validation tasks that are required for the corpus. During the validation stage the progress of the validation centre and its requests for additional information will be handled.

8. Conclusion

We have reported on ongoing work in building a balanced reference corpus of Dutch in the STEVIN project SoNaR. We have situated our work in the context of traditional written text corpora and in the context of web-as-corpus corpora. We have discussed pros and cons of both approaches. We have compared a number of corpora according to a range of important features. We conclude that the potential of SoNaR for future research and language technology developments fully warrants the effort undertaken.

9. Acknowledgements

The SoNaR project is funded by the Nederlandse Taalunie (NTU: Dutch Language Union) within the framework of the STEVIN programme under grant number STE07014. See also <http://taalunieversum.org/taal/technologie/stevin/>

10. References

M. Baroni and A. Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *EACL*

'06: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90, Morristown, NJ, USA. Association for Computational Linguistics.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

O. De Clercq and M. Montero Perez. 2010. Data Collection and IPR in Multilingual Parallel Corpora. Dutch Parallel Corpus. In *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC-2010)*, Valetta, Malta.

S. Evert. 2006. How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2):177 – 190.

S. Evert. 2008. A lightweight and efficient tool for cleaning web pages. In Nicoletta Calzolari et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.

M. Marx and A. Schuth. 2010. DutchParl. The Parliamentary Documents in Dutch. In *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC-2010)*, Valetta, Malta.

N. Oostdijk, M. Reynaert, P. Monachesi, G. Van Noord, R. Ordeman, I. Schuurman, and V. Vandeghinste. 2008. From D-Coi to SoNaR: a reference corpus for Dutch. In Nicoletta Calzolari et al., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.

M. Reynaert. 2008. Non-interactive OCR post-correction for giga-scale digitization projects. In A. Gelbukh, editor, *Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008. Lecture Notes in Computer Science Vol. 4919/2008*, pages 617–630, Berlin / Heidelberg. Springer.

C. Ringlstetter, K. Schulz, and S. Mihov. 2006. Orthographic errors in web pages: Toward cleaner web corpora. *Computational Linguistics*, 32(3):295–340.

I. Schuurman, V. Hoste, and P. Monachesi. 2010. Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. In *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC-2010)*, Valetta, Malta.