Beyond SoNaR: towards the facilitation of large corpus building efforts

Martin Reynaert[1], Ineke Schuurman[2], Véronique Hoste[3], Nelleke Oostdijk[4], Maarten van Gompel[4]

TiCC, Tilburg University[1]; CCL, Leuven University[2], UiL-OTS, Utrecht University[2]; University College Ghent/Ghent University[3]; CLST, Radboud University Nijmegen[4]

Abstract

In this paper we report on the experiences gained in the recent construction of the SoNaR corpus, a 500 MW reference corpus of contemporary, written Dutch. It shows what can realistically be done within the confines of a project setting where there are limitations to the duration in time as well to the budget, employing current state-of-the-art tools, standards and best practices. By doing so we aim to pass on insights that may be beneficial for anyone considering to undertake an effort towards building a large, varied yet balanced corpus for use by the wider research community. Various issues are discussed that come into play while compiling a large corpus, including approaches to acquiring texts, the arrangement of IPR, the choice of text formats, and steps to be taken in the preprocessing of data from widely different origins. We describe FoLiA, a new XML format geared at rich linguistic annotations. We also explain the rationale behind the investment in the high-quality semi-automatic enrichment of a relatively small (1 MW) subset with very rich syntactic and semantic annotations. Finally, we present some ideas about future developments and the direction corpus development may take, such as setting up an integrated work flow between web services and the potential role for ISOcat. We list tips for potential corpus builders, tricks they may want to try and further recommendations regarding technical developments future corpus builders may wish to hope for.

Keywords: Dutch Reference Corpus, corpus building, text conversion

1. Introduction

The construction of a large and richly annotated corpus of written Dutch was one of the priorities of the STEVIN programme conducted by the Dutch Language Union. Such a corpus, sampling texts from conventional and new media, is invaluable for scientific research and application development. The present paper briefly describes the Dutch reference corpus developed in the STEVIN-funded SoNaR project.[1] The construction of the corpus has been guided by (inter)national standards and best practices. Through the achievements and the experiences gained in the SoNaR project, a contribution was made to the further advancement of the standards and tools and the dissemination of the corpus. In the spirit of (Schuurman et al., 2004), we discuss what we see on the basis of experiences gained in building SoNaR, as desirable further developments which should help facilitate other future large-scale corpus building efforts.

The construction of a large reference corpus that aims to serve the interests of a wider research community requires that a number of prerequisites are met. Thus, in order to build a reference corpus for a language, one needs to incorporate more, and more diverse, text types than are freely accessible on the web and that are fully automatically processable. Furthermore, in order to build a balanced corpus for a language, one has to be able to selectively sample the texts to be incorporated. Finally, in order to build a widely available corpus one has to settle Intellectual Property Rights (IPR) to the fullest extent possible. Otherwise the fruits of one's labours are reduced by other researchers' limited access to the contents, reduced to access of mere snippets of texts or word frequency information only.

In what follows we outline tips, tricks and further recommendations towards possible future endeavours of this kind for other languages. We discuss the ramifications of the 'project' format under which we have worked. Inherent constraints of a 'project' are its limited duration in time and the confines of its budget. We highlight the major consequences of these constraints on the SoNaR corpus, as it is now. In the next section, we first describe what is now widely available through the Dutch-Flemish HLT Agency.[2]

2. The SoNaR reference corpus for contemporary, written Dutch

In this section we describe the design of the reference corpus and its contents and provide a brief description of linguistic annotations performed.

2.1. Corpus Design and Contents

The design of the reference corpus profited from the experiences in other large scale projects directed at the compilation of corpora (e.g. the British National Corpus (Aston and Burnard, 1998), the ANC (Ide et al., 2000) and the CGN (Schuurman et al., 2003)). In addition, consultation of the user community contributed to establishing needs and priorities.

The design was ambitious as it aimed at a 500 millions of word tokens (MW) reference corpus of contemporary standard written Dutch as encountered in texts (i.e. stretches of running discourse) originating from the Dutch speaking language area in Flanders and the Netherlands, as well as Dutch translations published in and targeted at this area. Texts were to be included from more conventional genres and text types as well as from the new media. The corpus was to include native speaker language and the language of (professional) translators. It was intended that approximately two-thirds of the texts would originate from the Netherlands and one-third from Flanders. Only texts were

[1] http://taalunieversum.org/taal/technologie/stevin/

[2] http://www.inl.nl/tst-centrale/nl/home

to be included that had appeared from the year 1954 onwards.[3]
In the course of the SoNaR project the corpus design originally conceived was modified.[4] There were several reasons for this. Thus, it proved impossible to maintain that texts which had not been printed but pre-electronically typed, were going to be incorporated, for we did not have the means to digitalize and properly correct these. In other cases a component (for example, web sites) would include rather different text types, each of which when examined more closely was considered to constitute a component in its own right (for example, blogs, discussion fora). Moreover, as we found the acquisition and preprocessing of certain types of data quite problematic we decided on more realistic targets (e.g. 0.5 MW of SMS instead of 5 MW). Finally, the enormous flight Twitter has taken, was a development we did not anticipate and was cause for modifying the design. In fact, the original design did not envisage the collection of tweets or even blogs at all.
We have created two corpora: the full 500 MW reference corpus[5] (further: SoNaR-500) and a richly annotated subset of 1 MW (SoNaR-1).

2.2. SoNaR-500: Corpus Format

The SoNaR-500 corpus is delivered in FoLiA XML format. FoLiA (short for "Format for Linguistic Annotation") (van Gompel, 2012) is an extensible XML-based annotation format for the representation of linguistically annotated language resources. It introduces a flexible paradigm independent of language, label set or linguistic theory. It is designed with the principles of expressivity, uniformity and extensibility in mind. Central to the paradigm is the notion of each annotation instance being of a certain class, i.e. the actual annotation value such as a particular part-of-speech tag. This class in turn pertains to a certain set; a label or tag set that either implicitly or explicitly defines all classes within its scope. Each type of annotation is consistently implemented as an XML element, which can furthermore always be enriched with one of several generic FoLiA XML attributes such as for example "annotator", representing the name or identifier of the person or system responsible for the annotation instance. A mixture of in-line and stand-off annotation is employed to deliver the expressive power necessary for the various annotation types. Except for data originating from the social media (chat, twitter, SMS), SoNAR-500 is annotated for parts of speech, lemmata, and named entities.
The aspiration of FoLiA is to be a universal but practical "one format fits all" framework, preventing users of having to cope with a wide variety of different formats for different annotation types, or having to invent ad-hoc extensions to a format whenever a resource is augmented with new annotation types. FoLiA takes an approach in which a single XML file represents an entire document with all its linguistic annotations. The format is intended to be used as a universal storage and exchange format for language resources, including corpora such as SoNaR-500.
Accompanying each FoLiA-formatted text in SoNaR is a CMDI-metadata file which contains the available associated metadata (Broeder et al., 2011).

2.3. SoNaR-1: Corpus Annotation

The **syntactic annotation** in SoNaR was inherited from another STEVIN project, called LASSY (van Noord et al., 2010). In this project a manually verified syntactically annotated 1-million word corpus was developed. This corpus served as the basis for four semantic annotation layers. These layers, which include the annotation of named entities, co-referential relations, semantic roles and spatio-temporal relations, were also completely manually checked. Where tools were available for pre-annotation, as was the case for semantic role labeling and spatio-temporal annotation, the task was redefined as a correction task. In case no tools were available (for Named Entity Labeling) or if the current tools were not considered performant enough (as for the annotation of coreferential relations), the annotation was done completely manually.
Through the annotation of **Named Entities**, we now have access to a balanced data set, which will allow for the creation and evaluation of supervised named entity recognizers for Dutch. The corpus covers a wide variety of text types and genres in order to allow for a more robust classifier (Desmet and Hoste, 2010), and better cross-corpus performance. We will discuss the granularity of the annotations and illustrate this by means of an example. Through the fine granularity of the labels (see example 1), it becomes possible not only to differentiate between the literal and metonymic use of named entities, but also to classify named entities in more detailed subtypes.

> *Example:* "We verwelkomen de betrokkenheid van [Rusland][LOC.country.meto.human] als partner en we hopen dat [Rusland][LOC.country.meto.human] ook aan het vervolg op [Kyoto][LOC.bc.meto.misc] zal deelnemen." (Eng: We welcome the participation of Russia as a partner and hope that Russia will also keep on participating after Kyoto.)

Through the annotation of the **co-reference relations**, we created one of the largest data sets currently available to coreference resolution research, which not only allows for the study of identity relations between nominal constituents, but which also has basic annotations for bridging references. Furthermore, the balanced nature of the data also allows for studying cross-genre performance (De Clercq et al., 2011).
The Annotation of **Semantic Roles** brought us the adaptation and extension of an existing set of guidelines to Dutch (Monachesi et al., 2007) and a Dutch version of the PropBank frame index. As for the other annotation layers, the data set will not only allow us to build an SRL system for

[3] In the year 1954 a major spelling reform was put into effect, as a result of which from this year onwards a common spelling of the Dutch language came into use in Belgium and the Netherlands.

[4] An overview of the original design can be found in (Oostdijk, 2006). In the report also the motivation for this design is given.

[5] The reader interested in an overview of the current contents of the corpus per text type and per country of origin will find the statistics at http://hmi.ewi.utwente.nl/sonar_language

Dutch, but also to assess its performance on a variety of text genres (De Clercq et al., 2012).

Recognition and normalization of **spatiotemporal expressions** (STEx) in SoNaR was done using a large database reflecting the spatiotemporal knowledge of people living in Belgium and The Netherlands, i.e. of the people who were meant to read texts as those contained in SoNaR.

The annotations for recognition and normalization of spatiotemporal expressions are rather finegrained in order to facilitate reasoning, especially on the basis of several documents,[6] at a later stage. Currently this level of annotation concentrates on geospatial and temporal expressions, in combination, i.e. there is one level of annotation instead of separate ones for spatial and for temporal annotation (Schuurman and Vandeghinste, 2010), (Schuurman and Vandeghinste, 2011).

2.4. Beyond SoNaR: TTNWW

Most of the steps involved in creating the SoNaR corpus as described in this paper are also necessary when other documents are to be analysed. Of course, only up till the level one is interested in. Within SoNaR, almost all levels were handled on their own, an automatic flow of information from one level of analysis to the other was out of question even when the output of level `n` did serve as input for level `n+1`.

This is remedied in the Flemish-Dutch CLARIN pilot project TTNWW (TST Tools voor het Nederlands als Webservices in een Workflow).[7] In this project the goal is to create an environment in which someone can use a web interface for a specific document to be analysed, for example with respect to Named Entities, or spatiotemporal relations, and at some moment in time she will get a mail announcing where the analysed document can be found.

Web services in the workflow are encouraged to employ an integral and expressive annotation format such as FoLiA (cf. section 2.2.) as data-exchange format, or use a more constrained task-specific format. Wrappers need to be written and included in the pipeline whenever format conversions are needed, and when integration of different layers of annotation has not yet taken place. An added complication to this integration process is that for some annotation types, the full text is represented in one, possibly huge, file (like PoS, STEx), whereas for example for syntactic analysis every sentence is represented as a separate file. Usage of FoLiA intends to alleviate such problems.

Another issue is the role of ISOcat, a linguistic concept database developed by ISO TC 37 to provide reference semantics for annotation schemata. Within TTNWW all metadata and annotation schemes used are defined in ISOcat in order to be able to decide what the relation is between a specific notion X (like *noun*) in layer A (e.g. PoS) and layer B (e.g. syntactic annotation).[8] Or the other way around, two notions A and B at several levels referring in se to the same linguistic notion C. In order to promote smooth cooperation between layers of annotation, such (non-)relationships are to be established. This can be done using ISOcat, especially when also the new extensions RELcat and SCHEMAcat are used to relate linguistic notions (DCs), even between languages, cf. (Schuurman and Windhouwer, 2011). ISOcat will also enable linking with other (de facto) standards and best practices. In order to promote this, existing definitions were reused whenever possible, i.e. when they are not in conflict[9] with the way they are used in SoNaR, and thus in TTNWW. And when new definitions were made, they are related (via RELcat) to existing ones, if any.

3. Tips towards the facilitation of future corpus building efforts

In this section we describe developments during SoNaR which have proved fruitful, to greater and, in some cases, to lesser extent.

3.1. Intellectual Property Rights

The reference corpus is intended to serve and be available to the wider research community. Therefore, considerable efforts were put into the settlement of the intellectual property rights (IPR). This was done in close collaboration with the Dutch HLT Agency who is responsible for the distribution of the corpus and its future maintenance. While the HLT Agency arranges the licenses with prospective end users (academics and other non-profit institutes but also commercial parties) before granting them access to the data, it was the responsibility of the corpus compilers to make sure that IPR was settled with the content owners who agreed to have their texts included in the corpus. To this end, the HLT Agency provided model contracts that the corpus compilers could use.

We started off with two types of contracts, one aimed at individual donators and one aimed at publishers. but gradually moved to less cumbersome ways of settling IPR. The original contracts were paper-bound, requiring the three parties involved (the donator, the HLT Agency acting for the Dutch Language Union and the University doing the acquisition) to sign and send back and forth the actual copies of the agreement. This works for donations involving substantial amounts of texts, but not for, say, an individual student's term paper, let alone for a single email. Table 1 lists the types of agreements we eventually used.

Eventually, every agreement is given a code, which encodes a lot of valuable metadata about the agreement/donation. This is a vital part of the logistics of corpus building, since it is absolutely necessary that each snippet of text can be traced back to its origin. For example: the license agreement code 'SoNaR.2BC.NL-B.00015' tells us that this is an agreement reached within SoNaR, that a standard agreement for publishers (2) has restrictions (B: e.g. no agreement concerning commercial use, i.e. texts were donated for research purposes only) and that other particulars (C: e.g. which subpart of the donator's website falls within the terms of the agreement) are described in the agreement itself. The code further indicates that this is an agreement

[6] Like multidocument summarization.

[7] In English: "HLT Tools for Dutch as Webservices in a Work Flow." This project runs till October 2012.

[8] There indeed are tokens considered a *noun* at the syntactic level, but not at the level of PoS tagging.

[9] Being too broad, too narrow, too vague, too language dependent, ...

Code	IPR-agreement	Type
1	Standard agreement for individuals	Paper copy in threefold, signed by all parties
2	Standard agreement for publishers	Paper copy in threefold, signed by all parties
3	Email agreement	Email received from donator saying we could use such and such texts
4	Explicit agreement	Texts are publicly available online and identified as such by the website of origin by e.g. an appropriate Creative Commons License or near-equivalent statement: ('Reproduction permitted with due acknowledgement.')
5	Online agreement	Sent to donator upon donation via the drop-box
6	Implicit agreement	Donator has used one of the online channels for donation of e.g. email, SMS
7	No agreement	Text widely available online, no possibility of reaching author(s) for settling IPR. Texts are often anonymous (e.g. SPAM)

Table 1: Types of IPR-agreements used in SoNaR

concerning texts from The Netherlands (NL), acquired by project partner (B) Tilburg University. It was the 15th (00015) agreement reached by this partner.

3.2. Acquisition

As we wanted the corpus to reflect the large degree of variation found not only between text types but also within one and the same text type, acquisition efforts were directed at including texts from a large variety of sources.[10] The identification of potential text providers was done on an ad hoc basis using various means available. Thus the networks of project members and associates were tapped into, contacts were established and major agreements arranged with television broadcasting companies, the conglomerate of national newspapers, major publishers of periodicals and other large text providers, while many other candidates were identified on the basis of their web presence and duly contacted. As a result of the attention the creation of the reference corpus attracted from the media, occasionally we would be approached by people offering data or giving pointers to interesting data sets.

Where we were aware of other text collections that held Dutch data representative of specific text types (such as JRC-Acquis for legal texts or the OPUS Corpus which includes Dutch subtitles), we have pursued the inclusion of these data.[11] This course of action was motivated by the idea that in the SoNaR project we would impact an added value in yielding the XML uniform to the other data in the reference corpus, but also through the tokenization and further linguistic annotations we provide automatically: POS-tagging, lemmatization and Named Entity labelling.

[10]It should be noted that on principle we never paid for the acquisition of data and the settlement of IPR. Sometimes we would pay a small fee for the extra work that a text provider put into delivering the texts in a form that for us was easier to handle. In the SMS campaign there was the chance of a prize for those who contributed data.

[11]JRC-Acquis is a collection of parallel texts from the EU comprising the contents, principles and political objectives of the Treaties; EU legislation; declarations and resolutions; international agreements; acts and common objectives (Steinberger et al. 2006). The OPUS Corpus is an open parallel corpus which is publicly available. See also http://opus.lingfil.uu.se/

3.3. Data Transfer

Arrangements must be made for the actual transfer of the acquired data. What is all too readily overlooked is that the ease with which data can be transferred from the text provider to the corpus compiler can be a decisive factor in the successful acquisition of texts. If transfer is complex and requires that effort be put into it on the part of the text provider, chances are that the provider will refrain from doing so.

There are various ways of making the transfer of data easy for data providers. One example is the use of a drop box. The SoNaR drop box[12] developed, unfortunately, rather late in the project has demonstrated its usefulness. It provided an easy interface to the text provider for uploading the (archives of) text files and for providing, at his/her own discretion some personal information for inclusion in the metadata. After submission, the text provider received a thank-you email which further contained the actual text of the IPR-agreement the text was subject to. Another example of how the transfer of data may be made easy is the way in which by means of an existing application SMS texts could be uploaded directly from Android mobile phones onto the SoNaR website[13].

At the beginning of this section it was observed that data acquisition was a formidable task. Indeed, identifying and acquiring the necessary data and arranging IPR for a corpus of 500 million words represented a major challenge. Yet, as such it is not so much the large quantity of data that one should be in awe of, it is the quantity combined with the diversity of text types that the corpus comprises that is truly ambitious. All through the project the balancedness of the corpus has been a concern. For some text types, we have reached a nice balance, for others this has remained problematic. Especially with texts directly obtained from the internet the amount of data tended to rapidly exceed the quantity envisaged in the corpus design. For example, the largest Flemish internet forum that we managed to arrange IPR with, by itself holds well over 500 million words of text. On the other hand, other text types were really hard to come by and were constantly at risk of being struck off the acquisition list. The corpus design was therefore used to control for balancedness and to ensure that apart from quantity there would be sufficient diversity: in a number of cases (such as the Flemish internet forum) only a fraction of the material is actual part of the 500 MW SoNaR corpus; the rest of the data is regarded as surplus. To the extent possible within the limitations of the project these data have been processed in the same manner and are available to those for whom there is never enough data.

[12]http://webservices.ticc.uvt.nl/sonar/

[13]The original application was developed by the National University of Singapore. It was adapted for use in the SoNaR project (Treurniet et al., 2012). Adaptation consisted primarily in translating the operating instructions for uploading SMS texts. Linked to this is a SoNaR website on which more information about the project and more instructions specific to different kinds of mobile (smart)phones could be found: http://www.sonarproject.nl/

3.4. Corpus (Pre)Processing

Various steps are to be taken in the preprocessing of the corpus, from the stage where texts have been acquired and delivered in their original formats, up to the point where they are available in a uniform XML format. The first step to be taken once the data had been acquired was to make the incoming data stream suitable for further upstream processing. It involved the conversion from the different file formats encountered such as PDF, MS-Word, HTML and XML to a uniform XML format. This uniform format should allow us to store metadata and the text itself along with linguistic annotations from later processing stages. Moreover, it provided the means to perform XML validation after each processing stage: first after the conversion from original file format to the target format, and then again whenever new annotations had been added. Especially the validation after the first conversion appeared to be a crucial one in order to prevent that the processing chain was jammed due to incorrect conversions.

Putting much effort in the development of conversion tools was regarded outside the scope of the project. However, the conversion from original format to target XML proved to be rather problematic in a substantial number of cases. Given the data quantities aimed at, an approach that uses a (semi-)manual format conversion procedure was not regarded a realistic option. Therefore the approach was to use existing conversion tools and repair conversion damage wherever possible. For a large proportion of the data this procedure worked quite well. Sometimes only minor adaptations to the post-processing tools were required in order to fix a validation problem for many files. Some parts of the collected data, however, had to be temporarily marked as unsuitable for further processing as it would take too much time to adapt the post-processing tools. Especially the conversion of the PDF formatted files appeared to be problematic. Publicly available tools such as pdf2html that allow for the conversion from PDF to some other format often have problems with columns, line-breaks, and headers and footers, producing output that is very hard to repair. On the other hand, as moving away from abundantly available content in PDF format would seriously limit the possibilities in finding a balance over text data types, the approach was to do PDF conversion semi-automatically for a small part of the collection.

While we have focused on PDFs in the foregoing, all other text formats have their own issues and varying amounts of effort are always required to convert other formats successfully to the target file format. Most time-demanding in this respect is the separation of metadata and text and the proper collection of both.

3.5. Lexical correction

In the SoNaR project, automatic lexical correction was performed on the frequency lists derived from the corpus, i.e. all divergent spelling variants were automatically lined up with their canonical form by means of TICCL (Text-Induced Corpus Clean-up), which was introduced in (Reynaert, 2008). In the course of the project we have continued to develop new approaches to large scale corpus clean-up on the lexical level. In (Reynaert, 2010) we report on a new approach to spelling correction which focuses not on finding possible spelling variants for one particular word, but rather on extracting all the word pairs from a corpus that display a particular difference in the bag of characters making up the words in the pairs. This is done exhaustively for all the possible character differences given a particular target edit distance, e.g. an edit distance of 2 edits means that there are about 120K possible differences or 'character confusions' to be examined to collect all variants within that limit for all words. This is cheaper than looking for all words' variants word by word.

3.6. Language recognition

Where deemed necessary or desirable during processing, we have applied the TextCat[14] tool for language recognition. Depending on the source and origin of the texts this was variously applied at document or paragraph level. Language recognition was never applied at sub-sentential level. Foreign language snippets or stretches of text abound in Dutch. The importance of properly distinguishing between what is Dutch and what is not is illustrated by the assertion in (Mihalcea and Nastase, 2002) that of the languages investigated on the basis of web corpora, Dutch has the most diacritics. We think this is due to the fact that Google at the time did not yet distinguish between Dutch and Frisian - the other officially recognized language in The Netherlands - and so the diacritics count for Dutch was amplified by that for Frisian. The SoNaR corpus should allow for a proper re-assessment of this probably incorrect assertions about the language's properties.

4. Tricks towards the facilitation of future corpus building efforts

In this section we will describe developments that are thought to be achievable within a project and are thought to be potentially highly fruitful for further corpus building efforts.

4.1. Perennial acquisition

In fact, the above describes an opportunistic approach to corpus building based on an ad hoc strategy imposed by the constraints of the project. Ideally, however, in building what amounts to be the major corpus representing the language of over 20 million people, the effort should not have been limited by a project's time of duration. We have reached agreements with major text providers which might continue long after SoNaR.

High-level IPR negotiations may facilitate corpus building on an ongoing, long term basis. This should be the norm, rather than the exception as it was in SoNaR. Ideally, one would have an infrastructure allowing for continuous and continuing donations. Some donators have sent us new material according to an agreed schedule. One major publisher of periodicals uploaded the XML-versions of each week's new editions. The translation office of a Belgian civil service sent monthly batches of their latest output.

Especially for collecting the amounts of word tokens foreseen in our corpus design for particular text types, another

[14] http://www.let.rug.nl/vannoord/TextCat/

strategy would be called for. For emails and student assignments, one would not chase individuals but squarely target the institutions where these are produced. This would entail, probably high-level, negotiations with the boards of directors e.g. of universities to have the institution's statutes changed in such a way that copyright on student assignments no longer falls to each individual student, but rather to the university. Which can then decide to make the year's crop available for scientific research through incorporation in the national corpus. A similar acquisition effort may then imply and effect that the organization's collective administrative, non-personal email output is also sent to the corpus email account.

The change one hopes for in e.g. a university's statutes involves copyright transfer from the individual students to their university. This is in no way what one should even mention in dealing with publishers. In our experience, it is generally unproductive to mention the word 'copyrights' unless one makes it very clear from the very onset that this is emphatically not what one is after. What one is after is not a change of hands of copyrights. One is after the right, bestowable only by the copyright holder, to have an electronic version of the book(s) and the right to distribute this, as part of a much larger text collection, to researchers in academia, possibly also in commercial organizations. In dealing with publishers, we have learned that all seem to have their own private ideas about or interpretations of what exactly is copyright and who is in fact a or the copyright holder and whether or not also permission should be sought from the translators and/or original book publishers and/or foreign language authors. Furthermore, individual publishers are not particularly keen on being approached with requests they may find hard to interpret. We have repeatedly been asked why we bother them directly, why we e.g. did not address the publishers' union or the national library. The thing here is that these would entail high-level negotiations probably requiring extended periods of time and most of us are just junior staff on temporary contracts trying to make the best of it. This job is probably best undertaken by people of renown, working in solid institutes of at least national fame who can bide their time and time the hands they play well.

4.2. Better tools

There is, alas, no panacea for text conversion today. While one expects to see the necessary artificial intelligence being gradually developed as the cultural transition from paper to digital copy proceeds, in SoNaR we have in part had to rely on human intelligence for properly extracting the running text from countless documents. We will further discuss which were the implications as for the moment there do not seem to be other solutions. An overall cure will probably require the combined efforts of page segmentation technology being developed mainly by researchers working on optical character recognition, i.e. text segmentation and page layout analysis specialists, and of Natural Language Processing specialists. For a recent appraisal of the state of the art in PDF text extraction technology we refer the interested reader to a recent technical paper released by Mitre (Herceg and Ball, 2011). Its main conclusion is that all too often valuable textual information is irretrievably lost when extracting text from PDF even when one uses the currently best-of-breed PDF text extractor available. In SoNaR, this has meant rejection of documents that did not meet the desired standard of text quality or, in cases where the text was deemed too valuable to let go, manual extraction of the text by laborious copying and pasting. The main obstacle to proper artificial intelligence-based text conversion is that most texts are unique in their overall layout and composition. More often than not, the texts within a batch donated even by a single provider do not form a single homogeneous collection as regards build-up. This is invariably the case with e.g. brochures, all of which come in their often often highly elaborate, conversion-unfriendly, layouts.

We have nevertheless expended effort in building an online tool aimed at potential volunteers to help with the formidable conversion task we had at hand. To this end a master student developed the web application WINKLE (Web Interface for Narrative Kernel Labeling and Extraction) (Persoons, 2010).

WINKLE first presents two automatically converted versions of the text chosen to the user. The two versions were produced by the Unix tools 'pdftotext' and 'pdftohtml'. The idea was that either of them might prove to have been converted more successfully than the other and to call on the user to decide on which to select for further processing. After this, the user would be presented with the original image of the text on the left of his screen and the flat text to be annotated to the right. Annotations effected were presented in colours visually denoting the type of annotation. We learned three main lessons from this exercise. First, the loss of visual information concerning the role of text segments contained in the original layout as a result of the automatic text conversion is a major hindrance and seriously raises the level of annotation difficulty. Second, the task of manually annotating flat text even with the original layout displayed alongside, especially in the absence of a guiding visual link between flat text and actual location of the text on the original text image, is too complex to explain briefly to potential volunteers, mainly due to the vagaries of automatic text extraction with the tools we used. Third, once we had the tool online, we realised that creating a community of volunteers around the effort would mean fostering this community. This was a responsibility we felt we could not take upon ourselves through lack of time and through the limited duration of the project as such. As a result, the tool never reached its full potential and was not widely advertised, nor indeed used. We nevertheless believe there might be a future in crowd-sourcing conversion tasks, given better online tools.

The situation today To conclude this section we will outline our vision for future corpus building tools. We will first highlight some aspects of what we see as lacking in SoNaR which should help clarify to potential users what to expect and what not to hope for in the present corpus.

First, all texts incorporated in the corpus have been converted to what is essentially a flat textual format. This offers no visual layout, text which in the original stood out either through font size, bold or italics mark-up or through any other possible means, is now reduced to a uniform

look. Especially poignant in the case of originally hyperlinked pages e.g. as on any web site, all possibly interesting links between pages have been lost. Probably especially for sociolinguistic research, the interaction between e.g. various posts on an internet forum is most probably no longer retraceable, except perhaps in the actual numbered sequence of the individual SoNaR documents representing all the various posts from particular internet fora incorporated in SoNaR. Especially towards writing studies, there is no record whatsoever of the actual writing or editing process the texts have undergone in their creation. In the SoNaR documents, there is no trace left of tables and figures in the original texts. Also, bibliographical references, and in most cases, footnotes and the captions of tables and figures have been summarily removed. The same holds for foreign language snippets of texts. The user should therefore not expect the running discourse that is available to be still fully interpretable or even intelligible in light of these excisions, which may affect not only human comprehension, but also some layers of annotation such as coreference or spatiotemporal analysis. Regarding metadata, more often than not what one would hope to have or find preserved, is not available.[15] Even major newspaper or periodical publishers who these days have external companies to manage and market their digital textual legacy often prove not to have preserved a record of e.g. the newspaper section the articles were published in.

What might be tomorrow We think that the creation of an add-on to available word processors could facilitate continuous corpus building through better handling of the text as well as the necessary metadata and through automatic incorporation into the corpus being built. Texts donated through the add-on via dedicated web services would be far richer than the final versions incorporated in SoNaR in that they would have retained the full edit history and may thus serve to study the writing process itself.

On the Corpora-List on 04/14/2011 in the thread 'Spell checker evaluation corpus', Stefan Bordag wrote: "perhaps producing such a corpus wouldn't be so difficult after all. Perhaps all it takes is a custom plugin for Open Office which people can use when they review documents they write in OO for errors. In this plugin, simply by clicking some accept button provided by the plugin they'd consent to have both the original version and the revised version sent to some database known to the plugin. With some time perhaps a sizable collection of all sorts of corrections in all sorts of languages could be produced by this.". What Bordag defines appears to be a killer application for corpus building in general. Setting up this kind of system implies that people donate their texts and their texts' editing history. The manner in which this is done would in fact allow for the fully automatic, community-driven building of corpora of contemporary written text. For any language, for any kind of corpus research.

This would solve the two major bottlenecks we have encountered daily in building SoNaR: IPR-settlement and metadata/text processing. Who better than the actual author of a text at time of donation to supply the necessary metadata? Metadata types one seeks to collect:

- personal: allowing the author to determine what level of personal information (s)he wishes to be associated with the particular text
- text: information about encoding, text type, register, style
- language: with possibility of indicating his/her level of proficiency
- processing: whether spelling/grammar checking was applied, using which particular tools...
- etc.

However casually they are mentioned, some types of information listed above have not and could not be collected for incorporation in SoNaR. In the proposed scheme, these metadata would automatically be incorporated in a suitable metadata scheme (e.g. CMDI) and the text itself, properly segmented in sections, paragraphs etc. with proper identification of headers/footers, tables, pictures, etc. saved in a suitable XML-format and sent on. This is a far cry from what one may currently hope to obtain from an automatic PDF-conversion and quite a step nearer the artificial intelligence we think will be required.

The receiving web service would then incorporate the text into the appropriate subcorpus according e.g. to text type, assign it the proper, normalized file name with the appropriate file number and further make it available to other web services for further linguistic enrichment: tokenization, POS tagging, automatic correction/normalization, syntactic parsing, etc. This would also entail gathering the immensely valuable information on the writing process itself, given the included edit histories.

5. Further Recommendations towards the facilitation of future corpus building

In this section we describe desirable developments that might well prove fruitful to large corpus building efforts, but require either an extended project format, or the widespread adoption of an altogether better text format.

5.1. Better Text Formats

Another idea, which would no doubt require a far larger and wider effort, would be to work towards developing better PDF text extraction facilities. We will describe how this may rely on taking into account the PDF generation method used as well as how embedding XML into PDF would provide a far more solid solution. As a first step, PDF convertors might be made sensitive to the PDF generation techniques as implemented in the various PDF generators. The Mitre technical paper clearly shows the huge effect the generator has on the extraction quality of the embedded text. We know of no convertors that take account of and exploit the PDF generation information available in the actual PDF files.

[15] For example, not knowing the day of publication of a newspaper makes it impossible to resolve an expression like *tomorrow* in an adequate way.

As a greater goal for the future, one might wish to broadly see adopted a PDF format which allows for the invisible embedding of XML information within the PDF document. Apparently, this has been possible since Acrobat version 4, released in April 1999. This would nevertheless need to be adopted by both PDF generator and extractor software producers, actively used by publishers and in time by the general public, to be of real use to corpus builders and have a significant impact on their efforts.

6. Concluding remarks

Through the SoNaR project two further important Dutch corpora have become available: the SoNaR-500 corpus and the SoNaR-1 corpus. In this paper we have shared our insights towards facilitating future large corpus building efforts.

7. Acknowledgements

The SoNaR project was funded by the Nederlandse Taalunie (Dutch Language Union) within the framework of the STEVIN programme under grant number STE07014. FoLiA was developed in the joint CLARIN-NL and CLARIN Flanders project TTNWW, financed by the Dutch and Flemish (via EWI) governments. TICCL is further being developed in NWO project Political Mashup.

8. References

G. Aston and L. Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

D. Broeder, O. Schonefeld, T. Trippel, D. Van Uytvanck, and A. Witt. 2011. A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

O. De Clercq, V. Hoste, and I. Hendrickx. 2011. Cross-Domain Dutch Coreference Resolution. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria. RANLP 2011.

O. De Clercq, V. Hoste, and P. Monachesi. 2012. Evaluating automatic cross-domain Dutch semantic role annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. LREC-2012.

B. Desmet and V. Hoste. 2010. Named Entity Recognition through Classifier Combination. In *Computational Linguistics in the Netherlands 2010: selected papers from the twentieth CLIN meeting*.

P. Herceg and C. Ball. 2011. A comparative study of PDF generation methods: Measuring loss of fidelity when converting Arabic and Persian MS Word files to PDF. Technical Report MTR110043, Mitre.

N. Ide, C. Macleod, C. Fillmore, and D. Jurafsky. 2000. The American National Corpus: An outline of the project. In *Proceedings of International Conference on Artificial and Computational Intelligence*. ACIDCA-2000.

R. Mihalcea and V. Nastase. 2002. Letter level learning for language independent diacritics restoration. In *Proceedings of CoNLL-2002*, pages 105–111. Taipei, Taiwan.

P. Monachesi, G. Stevens, and J. Trapman. 2007. Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop*, Prague, Czech Republic. ACL.

N. Oostdijk. 2006. A Reference Corpus of Written Dutch. Corpus Design. TR-D-COI-06f.

B. Persoons. 2010. Web Interface for Narrative Kernel Labeling and Extraction. Technical Report HAIT Master Thesis series nr. 10-007, Tilburg University, NL.

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*, volume 4919, pages 617–630, Berlin, Germany. Springer Verlag.

M. Reynaert. 2010. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187. 10.1007/s10032-010-0133-5.

I. Schuurman and V. Vandeghinste. 2010. Cultural Aspects of Spatiotemporal Analysis in Multilingual Applications. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

I. Schuurman and V. Vandeghinste. 2011. Spatiotemporal annotation: interaction between standards and other formats. In *Semantic Computing*, Palo Alto, California, USA. IEEE-ICSC Workshop on Semantic Annotation for Computational Linguistic Resources.

I. Schuurman and M. Windhouwer. 2011. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have to Offer? In *Proceedings of Supporting Digital Humanities (SDH 2011)*, Copenhagen, Denmark.

I. Schuurman, M. Schouppe, T. Van der Wouden, and H. Hoekstra. 2003. CGN, an annotated corpus of Spoken Dutch. In *Proceedings of the Fourth International Conference on Linguistically Interpreted Corpora*, pages 101–112, Budapest, Hungary. LINC-2003.

I. Schuurman, W. Goedertier, H. Hoekstra, N. Oostdijk, R. Piepenbrock, and M. Schouppe. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ... In *Proceedings of LREC'04, Volume I*, Lisbon, Portugal. ELRA.

M. Treurniet, O. De Clercq, N. Oostdijk, and H. Van den Heuvel. 2012. Collecting a Corpus of Dutch SMS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. LREC-2012.

M. van Gompel. 2012. FoLiA: Format for Linguistic Annotation. Documentation. Technical Report 12-03, ILK Technical Report, TiCC – Tilburg University.

G. van Noord, I. Schuurman, and G. Bouma. 2010. Lassy Syntactische Annotatie. Technical Report 19455, University of Groningen.