Chapter 13
The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman

13.1 Introduction

Around the turn of the century the Dutch language Union commissioned a survey that aimed to take stock of the availability of basic language resources for the Dutch language. Daelemans and Strik [5] found that Dutch, compared to other languages, was lagging behind. While the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; [25]) addressed the need for spoken language data, the dire need for a large corpus of written Dutch persisted and the construction of a multi-purpose reference corpus tailored to the needs of the scientific research as well as commercial development communities was identified as a top priority in the creation of an infrastructure for R&D in Dutch HLT.

The reference corpus, it was envisaged, should be a well-structured, balanced collection of texts tailored to the 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 were to be 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

N. Oostdijk (✉)
Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: n.oostdijk@let.ru.nl

M. Reynaert
Tilburg University, Tilburg, The Netherlands
e-mail: reynaert@uvt.nl

V. Hoste
University College Ghent and Ghent University, Ghent, Belgium
e-mail: veronique.hoste@hogent.be

I. Schuurman
KU Leuven, Leuven, Belgium
e-mail: ineke.schuurman@ccl.kuleuven.be

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_13, © The Author(s) 2013

automatic abstracting that are based on underlying corpus-based techniques were expected to benefit from the large-scale analysis of particular features in the corpus. Apart from supporting corpus-based modeling, the corpus was to constitute a test bed for evaluating applications, whether or not these applications are corpus-based.

On the surface, all stakeholders agree that a large reference corpus of written Dutch would be invaluable for linguistic research and the development of profitable services that require advanced language technology. However, as soon as one starts making preparations for the collection of the text, and the definition of the minimal set of meta-data and annotation layers, it appears that different purposes may very well translate into very different requirements. A very large, balanced, richly annotated multi-purpose reference corpus is very different from the task-specific corpora that have been built in – for example – the DARPA programmes and the European CLEF programme. What is more, while some of the stakeholders (e.g. linguists, application developers and system integrators) may be able to formulate requirements and desires in the terms of their own disciplines and business areas, it is not straightforward to translate these formulations into technical requirements for a reference corpus. This is one of the reasons why in 2005 the STEVIN Dutch Language Corpus Initiative (D-Coi) project was initiated.

Although there were as yet no examples of the type of reference corpus aimed at, it was, of course, possible to derive boundary conditions from experiences with existing corpora and the major trends in the development of linguistics and language technology.[1] Thus, a modern reference corpus should not only sample texts from conventional media such as books and newspapers, but also from electronic media, such as web pages, chat boxes, email, etc. It was evident that inclusion of texts from these sources would pose (new) problems related to IPR, and that they would require the development of novel tools for the detection and annotation of typos, non-words, and similar phenomena that are less prominent in well-edited texts from the conventional printed media.

The D-Coi project was a pilot project that aimed to produce a blueprint for the construction of a 500-million-word (500 MW) reference corpus of written Dutch. This entailed the design of the corpus and the development (or adaptation) of protocols, procedures and tools that are needed for sampling data, cleaning up, converting file formats, marking up, annotating, post-editing, and validating the data.[2] In order to support these developments a 50 MW pilot corpus was compiled, parts of which were enriched with linguistic annotations. The pilot corpus should demonstrate the feasibility of the approach. It provided the necessary testing

[1] At the time (i.e. in 2004, at the start of the STEVIN programme) the American National Corpus (ANC; [16]) was probably closest to what was envisaged for the Dutch reference corpus as it also intended to include data from electronic media.

[2] Already in the planning phase, we realised the importance of adhering to (inter)national standards and best practices. Subsequently, wherever possible we have tried to relate to and build upon (the results of) other projects as well as re-use of resources and tools. Especially the CGN project has been particularly influential.

ground on the basis of which feedback could be obtained about the adequacy and practicability of the procedures for acquiring material and handling IPR, as well as of various annotation schemes and procedures, and the level of success with which tools can be applied. Moreover, it served to establish the usefulness of this type of resource and annotations for different types of HLT research and the development of applications.

There can be no doubt that as preparatory project the D-Coi project has been very useful. It provided the opportunity to come up with a design for a reference corpus in close consultation with the user community. Moreover, the compilation of the pilot corpus gave us hands-on experience with the work ahead of us, some facets of which we had underestimated before. With the insights gained we got a better view of what realistically could be done and what not. This has definitely proven to be advantageous as we were much better prepared when in 2008 we undertook the actual construction of the full reference corpus in the SoNaR project.[3]

In what follows we describe the various phases in the construction of the reference corpus. In Sect. 13.2 different aspects related to corpus design and data acquisition are discussed. Section 13.3 focuses on corpus (pre)processing, paying attention to the steps taken to handle various text formats and arrive at a standard XML version. Section 13.4 describes the various types of annotation and how they came about. Finally, Sect. 13.5 concludes this chapter.

13.2 Corpus Design and Data Acquisition

In this section we describe the design of the written Dutch reference corpus and its implementation, relating the strategies adopted in collecting different text types (including a wide range of texts from both traditional and new media) and the experiences in the acquisition and arrangement of IPR.

13.2.1 Corpus Design

The Dutch reference corpus was intended to serve as a general reference for studies involving language and language use. The corpus should provide a balanced account of the standard language and the variation that occurs within it. In doing so, it allows researchers investigating language use in a particular domain (e.g. medicine) or register (e.g. academic writing), or by a specific group (e.g. professional translators)

[3] The acronym SoNaR stands for STEVIN Nederlandstalig Referentiecorpus, i.e. STEVIN Dutch Reference Corpus.

to relate their data and findings to the general reference corpus. The corpus was also intended to play a role in the benchmarking of tools and annotations.[4]

The design of the Dutch reference corpus profited from the experiences in other large scale projects directed at the compilation of corpora (e.g. the British National Corpus, BNC – [1], the ANC and the CGN). In addition, consultation of the user community contributed to establishing needs and priorities.

The user requirements study [28] constituted a crucial step in the process of designing a Dutch reference corpus. The inventory of the needs and desires of linguists and members of the Dutch HLT community made by means of a web questionnaire, followed by consultation of the different user communities in focus groups, helped us decide on the priorities that should be set. Through the involvement of (potential) future users in this early stage we expected to avoid oversights and shortcomings that could easily result from too narrow a view on design issues and a limited awareness of existing needs. Equally important, user involvement throughout the design stages of corpus creation would contribute to generate the necessary support for such an undertaking and knowledge transfer.

The design was ambitious as it aimed at a 500 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 to be included that had appeared from the year 1954 onwards.[5]

The design envisaged the inclusion of texts written to be read as well as texts written to be spoken, published and unpublished texts, and also of texts that had appeared in print or in electronic form, or had been typed (cf. Table 13.1). As we aimed for a balanced, multi-purpose corpus, the corpus was to include a wide range of text types, from books, magazines and periodicals to brochures, manuals and theses, and from websites and press releases to SMS messages and chats. Moreover, the sheer size of the corpus made it possible to aim for the inclusion of full texts rather than text samples, leaving it to future users of the corpus to decide whether to use a text in its entirety or to use only a select part of it that meets the sampling criteria that follow more directly from a specific research question.

In the specification of the design of the Dutch reference corpus we intentionally deviated from other previous corpus designs for reference corpora such as the BNC

[4] Cf. the definition of a reference corpus provided by EAGLES: "*A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials.*"

[5] 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.

Table 13.1 Overall corpus design in terms of three main design criteria, viz. intended delivery of the texts included, whether they were published or not, and the primary mode (electronic, printed or typed)

Written to be read 492.5 MW	Published 362.5 MW	Electronic 177.5 MW
		Printed 185.0 MW
	Unpublished 130.0 MW	Electronic 100.0 MW
		Printed 10.0 MW
		Typed 20.0 MW
Written to be spoken 7.5 MW	Unpublished 7.5 MW	Electronic 2.5 MW
		Typed 5.0 MW

and ANC. Especially the inclusion of much larger volumes of electronic texts, both published and unpublished, caused experts from the Center for Sprogteknology (CST, Copenhagen) charged with the evaluation of the design to raise questions as to its justification. Concerns were voiced as regards the effect the inclusion of such high quantities of electronic text would have on corpus quality, the arrangement of IPR, and thus on the representativeness and the balance of the corpus. At the same time the experts were receptive to the idea of an alternative design as they could well imagine that

> "the corpus will be contributing to, or may even be setting future best practices with regard to the proportional representation of electronic publications in reference corpora, because the existing guidelines that can be derived from the current large reference corpora, BNC and ANC, may need some additions. Text types like e-mail and discussion lists, chat and SMS are highly influenced by the intentions of quick, personal communication and by the requirements/limitations of the medium of communication as regards their functional style and language which differentiate them from traditional written text types. However, the need for novel NLP tools appropriate for new communication channels such as web chats, blogs, etc. justifies the high inclusion rate of such text types in a corpus intended to serve as a linguistic resource for the development of such NLP methods and tools." [2, p. 7]

In the course of the SoNaR project the corpus design originally conceived in the D-Coi project was modified.[6] There were several reasons for this. As we found that preprocessing typed texts was very laborious, time-consuming and error-prone, we decided to refrain from including this type of material. In other cases, such as

[6] An overview of the original design can be found in Table A.1 in the Appendix. For a detailed description and motivation we refer to [27].

with SMS messages where we found that the acquisition was quite problematic we decided on more realistic targets (e.g. 50,000 SMS texts instead of 5 MW).[7] 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 at all.

13.2.2 IPR

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 licences 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.[8] To this end, the HLT Agency provided model contracts that the corpus compilers could use.

IPR had to be arranged for texts from all kinds of sources, both in the public but also in the more private domain. With texts from the conventional printed media (such as books, magazines, newspapers) the situation as regards IPR is fairly clear.[9] IPR can usually be settled through the publisher. For texts that are born-digital and are apparently freely available on the internet (such as websites and discussion fora) arranging IPR, we found, is rather more tricky. In some cases IPR lies with the site owner as contributors at some point have consented to have their rights carried over. However, in many such cases it is unclear whether the data may be passed on to a third party. In other cases no apparent IPR arrangements have been made. As a result the IPR status of these data remains unclear and the rights probably remain with the original authors/contributors. With data from for example chat and SMS individual people must give their consent. It is especially with these more private types of data that people were hesitant to have their texts included in a corpus. Anonymisation of the data was considered but not further pursued as this would involve a great deal of work, while it would seriously impact on the authenticity of the data.

[7] cf. Sect. 13.2.3 and 13.3.1.

[8] 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.

[9] Although things may be complicated when texts have been digitised and placed on the internet (as for example those in DBNL – Digitale Bibliotheek Nederlandse Letteren, http://www.dbnl.org/).

In a number of cases there was no need to follow up on IPR matters as the texts were already available under some kind of licence, such as GNU GPL or Creative Commons, or by arrangement of law (the public's right to information).

13.2.3 Acquisition

Data acquisition has proven to be quite a formidable task. Ideally acquisition would be directed at texts that are already available in an open digital format so that the amount of work that must be put into making the text accessible can be reduced to a minimum. In actual practice we found that if we were to restrict the selection of data in this fashion this would seriously affect the balancedness of the corpus, especially since even with major publishers today the bulk of their holdings are not in (an open) digital format. In the acquisition process the primary aim was to identify and acquire texts that would fit the corpus design. And although we maintained a preference for formats that were readily accessible, we did not shy away from texts in formats that we knew would require considerable effort to preprocess.

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. The identification of potential text providers was done on an ad hoc basis using various means available to us. 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. 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.[10] 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 tokenisation and further linguistic annotations we provide.

For successful acquisition we found there is no single standard recipe. Different types of text and text providers require different approaches. Moreover, there are cultural differences: where potential text providers in Flanders may be persuaded to donate their texts arguing that the future of the Dutch language is under threat, in the

[10] 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" [44]. The OPUS Corpus is an open parallel corpus which is publicly available. See also http://opus.lingfil.uu.se/.

Netherlands the fact that by donating texts a contribution is made to science is what is found particularly appealing. The strategies used and experiences gained in the SoNaR project in approaching potential text providers, negotiating and successfully settling IPR have been documented in [8].[11]

Of course at some point 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. Although the SoNaR drop box we had at our disposal was introduced rather late in the project it has demonstrated its usefulness.[12] 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 represents 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. 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

[11] For the acquisition of tweets and SMS, special campaigns were organised (see [35, 47]).

[12] URL: 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. 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 (URL: http://www.sonarproject.nl/).

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.

Apart from having the data in the corpus represent various text types and topic domains, we also wanted the corpus to include both data originating from Flanders and data from the Netherlands. In a number of cases, as for example with the data from Wikipedia or JRC-Acquis, it was impossible to establish the origin.

All the text data files that were collected were gathered centrally and stored along with available metadata (such as content provider, date downloaded, original filename). An overview of the composition of the reference corpus can be found in Table A.1 in the Appendix.

13.2.4 Pilot Corpus

For the pilot corpus no separate design was made. In fact, the compilation of the pilot corpus ran very much in parallel to the work done in relation to the design of the 500 MW corpus and the development of procedures and the drafting of contracts that could be used for settling IPR matters. Given the primary aim of the pilot corpus, the main concern was that the corpus should be varied enough to be able to test the various procedures and protocols so as to avoid any omissions or oversights that might affect the compilation of the reference corpus.

In the compilation of the D-Coi pilot corpus, we found that IPR issues frustrated the acquisition process. In order to make sure that sufficient material would be available we therefore resorted to a more opportunistic approach of acquiring data. This involved focusing on data that were already available in the public domain (e.g. under a GPL or Creative Commons licence) or considered low-risk, such as texts found on public websites maintained by the government and public services.[14] Some genres and text types, however, remain underrepresented in the pilot corpus or do not occur in it at all. The latter is true for example for chat, email and SMS. Moreover, the corpus comprises relatively few Flemish data. An overview of the composition of the pilot corpus can be found in Table A.1 in the Appendix. The pilot corpus is described in more detail in [26].

13.3 Corpus (Pre)Processing

In this section we describe the various steps 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.

[14] 'Low-risk' meaning that remaining IPR issues could be expected to be resolved in the not too distant future.

13.3.1 Text Conversion

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.[15] 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 appeared 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. A varying amount of effort was required to convert other formats successfully to the target file format.

Progress of the work could be monitored by all project partners via a simple PHP web-interface[16] on a MYSQL database containing the relevant information for each file such as the raw word counts, validation status for each level, and total word counts (grand total, counts per document group, validated, etc.). The database was synchronised with the information in the D-Coi/SoNaR file system so that project partners could immediately fetch data that became available for their processing stage. The database and web-interface served as intermediate documentation of the work done.

[15] In the D-Coi project the XML format previously used in the CGN project was adopted with some slight changes. In SoNaR the D-Coi XML format was again modified (cf. also Sect. 13.5).

[16] URL: http://hmi.ewi.utwente.nl/searchd-coi

13.3.2 Text Tokenisation and Sentence Splitting

A major aim of the first conversion step to XML was to have titles and paragraphs identified as such. This is because most tokenisers, our own included, may fail to properly recognise titles and because the sentence splitting process expects a paragraph to consist of at least one full sentence. Failure in the first conversion step to recognise that a paragraph in TXT format is split up into n lines by newline characters, results in n XML paragraphs being defined. This is unrecoverable to the tokeniser. This fact can mostly be detected by the ratio of sentences identified after tokenisation in comparison to the number of paragraphs in the non-tokenised version. In such cases both unsuccessful versions were discarded and new ones produced semi-automatically by means of minimal, manual pre-annotation of the raw TXT version of the documents.

The rule-based tokeniser used was developed at the Induction of Linguistic Knowledge research team at Tilburg University prior to the D-Coi project. It was slightly adapted to the needs of the D-Coi/SoNaR projects on the basis of evaluations conducted by means of TOKEVAL, a tokeniser evaluator developed during the project in order to evaluate the available sentence splitters and tokenisers.[17] A very good alternative to the ILK tokeniser (ILKTOK), is the tokeniser that is available in the Alpino Parser distribution. As neither of the sentence-splitters/tokenisers available to us handled XML, we developed a wrapper program (WRAPDCOITOK) that deals with the incoming XML stream, sends the actual text to the sentence splitter/tokeniser, receives the outcoming sentences and tokens and wraps them in the appropriate XML. This scheme further allows for collecting sentence and word type statistics and for word type normalisation during the tokenisation step.

13.3.3 Text Normalisation and Correction

During the D-Coi project we developed CICCL, which is a set of programs for identifying various types of primarily typographical errors in a large corpus. CICCL stands for 'Corpus-Induced Corpus Clean-up' and has in part been described in [32]. Assumptions underlying this work are: (1) that no resources other than corpus-derived n-gram lists are available, (2) that the task can be performed on the basis of these resources only, to a satisfactory degree, (3) that in order to show that this is so, one needs to measure not only the system's accuracy in retrieving non-word variations for any given valid word in the language, but also its capabilities of distinguishing between what is most likely a valid word and what is not.

[17] These and other tools developed in the D-Coi project are available from http://ilk.uvt.nl, asareourtechnicalreports.

Where diacritics are missing and the word form without diacritics is not a valid word in its own right, fully automatic replacement was mostly possible and has been effected. This was performed for the words requiring diacritics which are listed in the [57], i.e. the official 'Word list of the Dutch Language'. Also we have a list of about 16,500 known typos for Dutch and most of the selections have been screened for these.

In the SoNaR project, text correction was performed more thoroughly, 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 [33]. In the course of the project we have continued to develop new approaches to large scale corpus clean-up on the lexical level. In [34] 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 what we call character confusions to be examined.

13.3.4 Language Recognition

Where deemed necessary or desirable during processing, we have applied the TextCat tool for language recognition.[18] 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. However, in the Wikipedia texts, paragraphs containing foreign UTF-8 characters above a certain threshold were summarily removed, not on the basis of a TextCat classification but on encoding alone.

For some batches, notably the posts from a Flemish internet forum primarily dedicated to popular music and thus mainly to adolescents, TextCat was used to classify all posts separately. We found that over half received the following TextCat verdict: "I do not know this language". The language in question almost infallibly being a dialectical variety of the poster's specific internet idiolect. These posts were included and their TextCat categorisation was included in the metadata.

13.4 Corpus Annotation

This section describes the various types of annotations that were added to either the full reference corpus (the SoNaR-500 corpus for short), or one of two subsets: the D-Coi pilot corpus or a set of one million words (the SoNaR-1 corpus for short, cf.

[18] TextCat is available from http://www.let.rug.nl/vannoord/TextCat/

Table 13.2 Composition of the SoNaR-1 corpus. In all SoNaR-1 comprises 1,000,437 words

Text type	# words	Text type	# words
Administrative texts	28,951	Manuals	5,698
Autocues	184,880	Newsletters	5,808
Brochures	67,095	Newspapers	37,241
E-magazines and e-newsletters	12,769	Policy documents	30,021
External communication	56,287	Press releases	15,015
Instructive texts	28,871	Proceedings	6,982
Journalistic texts	81,682	Reports	20,662
Legal texts	6,468	Websites	32,222
Magazines	117,244	Wikipedia	260,533

Table 13.2). A decisive factor as regards what annotations were added to which dataset was the availability of tools that were sufficiently mature to allow large scale, fully automatic annotation. For part of speech tagging and lemmatisation, and named entity recognition this is (now) the case. For syntactic and semantic annotation, however, the annotation process is at best semi-automatic (that is, when aiming for annotations of high quality).

Since it is generally believed that the lack of a syntactically and semantically annotated corpus of reasonable size (min. 1 MW) is a major impediment for the development of academic and commercial tools for natural language processing applied to the Dutch language, we invested in these types of annotations. The SoNaR-1 corpus was syntactically annotated and manually verified in the Lassy project while in the SoNaR project four semantic annotation layers were added. These layers, which include the annotation of named entities, co-referential relations, semantic roles and spatio-temporal relations, were completely manually checked. Where tools were available for pre-annotation, the task was redefined as a correction task.

13.4.1 Part-of-Speech Tagging and Lemmatisation

For the tagging and lemmatisation of the reference corpus we aimed to yield annotations that were compatible to those in the CGN project. To the extent possible we wanted to re-use the tag set as well as the annotation tools and protocols for the human annotators. The tag set used to tag the reference corpus is essentially the same as that used for the Spoken Dutch Corpus (CGN), be it that a few tags were added to handle phenomena that do not occur in spoken language such as abbreviations and symbols [50]. Moreover, some tags that already existed in the original CGN tag set in the D-Coi/SoNaR version cover additional phenomena.

In the D-Coi project the CGN tagger/lemmatiser was adapted and retrained so that it would be able to cope with written text. This new version of the tagger/lemmatiser, which went by the name of Tadpole, was used to tag and lemmatise the entire D-Coi pilot corpus.[19] PoS tagging with Tadpole reached an accuracy of 96.5 % correct tags (98.6 % correct on main tag) on unseen text.

For part of the pilot corpus (500,000 words) the tagging output of Tadpole was manually verified.[20] This was done with the idea that it would provide us with a qualitative analysis of its strengths and weaknesses, something we thought was of particular importance since the tagging-lemmatisation of the reference corpus would be done fully automatically (the sheer size of the corpus prohibited manual verification).

The task of manually verifying the tags was a bit of a challenge: the high accuracy output attained by Tadpole made it hard to find the few mistakes left, especially when looking through the tags one by one. We therefore deployed a tool that focused on suspect tags only (identified by a low confidence value).

The output of the tagger consisted of PoS tagged files, containing all possible tags for each token, together with the probability of that tag. We developed a tool for the manual correction of these automatically generated PoS tagged files. This tool takes a PoS tagged file as input, together with a threshold value. It presents the human annotator only with those cases where more than one possible tag has an above-threshold probability. All other cases where more than one tag is generated by the tagger, or those cases where only one tag is generated, are not presented to the annotator, resulting in a markedly lower workload.

We performed a small experiment to determine at which value we best set the threshold: a threshold value of 0.06 results in a reduction of the number of decisions to be made by the human annotator with 28 %, while skipping a mere 1 % of errors which are not presented to the annotator. This shows that, with the benefit of a tagger well-trained on a large volume of manually checked training material, we can manually check much larger amounts of data in the same time, missing hardly any errors. While following this procedure, all manually corrected material is regularly checked against a blacklist of typical errors made by the tagger, particularly on multi-word named entities and high-frequency ambiguous function words such as *dat* ('that', having the same ambiguity as in English) which the tagger sometimes tags incorrectly but with high confidence.

Except for some types of data originating from the new media, the reference corpus was tagged and lemmatised automatically using Tadpole's successor FROG.[21] In view of the huge amount of data and the high quality of FROG's output we refrained from any manual verification of the tagger-lemmatiser output. However,

[19] Tadpole is described in more detail in [49]. A more detailed account of how tagging and lemmatisation was actually applied in the case of the D-Coi pilot corpus is given in [48].

[20] At a later stage, another 500,000 words from the SoNaR corpus were manually corrected in the Lassy project. The total set of one million words is what we have elsewhere referred to as the SoNaR-1 corpus (cf. Sect. 3.4).

[21] FROG is available under GPL (online demo: http://ilk.uvt.nl/cgntagger/,software:http://ilk.uvt.nl/frog/). We refrained from applying FROG to data such as chats, tweets and SMS as we expected that FROG would perform very poorly on this type of data.

Table 13.3 Accuracy of Alpino on the manually corrected syntactically annotated part of D-Coi. The table lists the number of sentences, mean sentence length (in tokens), and F-score in terms of named dependencies

Corpus	Sentences	Length	F-score (%)
D-Coi	12,390	16	86.72

with the tool and procedure developed to support the manual verification of the data, users can yet undertake this task for specific subsets of the data as they see fit.

13.4.2 Syntactic Annotation

In the D-Coi project we also investigated the feasibility of (semi-)automatically annotating the corpus for syntactic information with Alpino, a computational analyzer of Dutch which was developed at the University of Groningen. Experiences with syntactic annotation in the Spoken Dutch Corpus (CGN) project had shown that the approach taken there was quite labour-intensive. Of course at the time of the CGN project, no syntactically annotated corpus of Dutch was available to train a statistical parser on, nor an adequate parser for Dutch.[22] However, at the start of the D-Coi project Alpino had sufficiently matured and became an option that deserved serious consideration while contemplating the syntactic annotation of large quantities of data.

Alpino provides full accurate parsing of unrestricted text and incorporates both knowledge-based techniques, such as a HPSG grammar and lexicon which are both organised as inheritance networks, as well as corpus-based techniques, for instance for training its disambiguation component. An overview of Alpino is given in [52]. Although the syntactic annotation scheme used by Alpino was based on the annotation guidelines that were developed earlier for the annotation of the Spoken Dutch Corpus, the annotation scheme deployed in D-Coi was not exactly the same as the one used in for the CGN [14, 42]. Differences include, for instance, the annotation of subjects of the embedded verb in auxiliary, modal and control structures, and the annotation of the direct object of the embedded verb in passive constructions. In the CGN scheme, these are not expressed. In D-Coi these subject relations are encoded explicitly.

Part of the pilot corpus (some 200,000 words) was annotated syntactically by means of Alpino and the annotations were manually corrected. In Table 13.3 we list the accuracy of Alpino on these data. With the syntactic annotations obtained by means of Alpino, we also inherited an XML format in which the syntactic

[22] An adequate parser should meet several requirements: it should have wide coverage, produce theory-neutral output, and provide access to both functional and categorial information.

annotations are stored. This format directly allows for the use of full XPath and/or Xquery search queries. As a result standard tools can be used for the exploitation of the syntactic annotations, and there is no need to dedicate resources to the development of specialised query languages.

After the D-Coi project was finished, syntactic annotation was further pursued in the STEVIN Lassy project. In this project, the one-million-word SoNaR-1 corpus was enriched with syntactic information. For more information we refer to Chap. 9, p. 147.

13.4.3 Annotation of Named Entities

Despite its huge application potential, the annotation of named entities and the development of named entity recognition (NER) systems is an under-researched area for Dutch. NER, the task of automatically identifying and classifying names in texts, has started as an information subtask in the framework of the MUC conferences, but has also been proven to be essential for information retrieval, question answering, co-reference resolution, etc.

The goal in the SoNaR project was to create a balanced data set labeled with named entity information, which would allow for the creation and evaluation of supervised machine learning named entity recognisers. The labeled data set substantially differs from the CoNLL-2002 shared task [45] data set, containing 309,686 tokens from four editions of the Belgian newspaper "De Morgen". First of all, the goal was to cover a wide variety of text types and genres in order to allow for a more robust classifier and better cross-corpus performance. Furthermore, instead of focusing on four named entity categories ("person", "location", "organisation" and "miscellaneous"), we aimed at a finer granularity of the named entities and we also wanted to differentiate between the literal and metonymic use of the entities. For the development of the guidelines, we took into account the annotation schemes developed in the ACE [11] and MUC (e.g. [4]) programmes, and the work on metonymy from [21]. In the resulting annotation guidelines, we focused on the delimitation of the named entities, after which each entity was potentially annotated with four annotation layers, covering its main type, subtype, usage and (in case of metonymic usage) its metonymic role (cf. Fig. 13.1).

The examples below clearly show that all tags maximally consist of four parts, in which the first part of the tag denotes the main type of the NE, the second part the sub type, the third one the use, and the last one the type of use.

1. Nederland[LOC.land.meto.human] gaat de bestrijding van het terrorisme anders en krachtiger aanpakken. Minister Donner[PER.lit] van justitie krijgt verregaande bevoegdheden in die strijd.
   (English: The Netherlands are planning to organise the fight against terrorism in a different and more powerful way. Minister of Justice Donner was given far-reaching powers in that battle.)

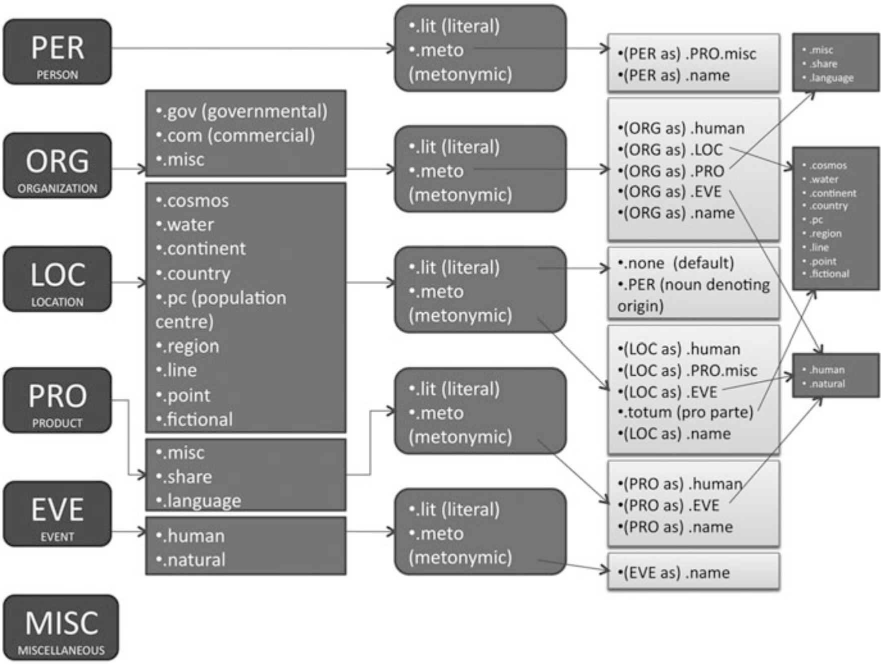

Fig. 13.1 Schematic overview of the named entity layers and the corresponding labels

2. Het is een eer om hier te zijn op MGIMO[ORG.misc.meto.loc]. Deze prachtige universiteit is een kweekvijver voor diplomatiek talent. Deze instelling heeft hechte contacten met Nederland[LOC.land.meto.human].
   (English: It is an honour to be here at MGIMO. This wonderful university is a breeding ground for diplomatic talent. This institution has tight connections with the Netherlands.)

The named entity annotations were performed on raw text and were done in the MMAX2[23] annotation environment. Annotation speed averaged around 3,500 words per hour. Taking into account the verification of the annotations by a second annotator, the actual annotation speed was about 2,000 words per hour. In order to evaluate the annotation guidelines, two annotators labeled eight randomly selected texts from the corpus (14,244 tokens in total). The interannotator agreement was measured with two evaluation metrics, namely Kappa [3] and F-measure ($\beta = 1$) [54]. The latter scores were calculated by taking one annotator as gold standard. The scores were calculated on five levels: span, main type, subtype, usage and metonymic role. For each level, scores were calculated on the entire set, and on a subset containing only those tokens on which both annotators agreed on the

[23] URL: http://mmax2.sourceforge.net

preceding level. For each of the levels, high agreement scores were obtained, with a Kappa score ranging from 0.97 to 0.91 and an F-score ranging from 99.6 to 98.9 %. For a detailed description of the guidelines and the interannotator agreement on each of the annotation levels, we refer to [10].

The annotated corpus was used for the development of a NE classifier [10], which was used for the automatic annotation of the remaining 499 million words. Although the one-million-word corpus already covered different text types, thus allowing to have a more balanced view on the quality of the named entity recogniser, this does not guarantee that the automatic labeling of the 499 million remaining words reaches the same accuracy levels. We expect that an adaptation of the classifier to informal text types (blogs, chats, sms) will be required. In order to allow for this adaptation, the full named entity recogniser was also delivered together with the manually verified annotations.

13.4.4 Annotation of Co-reference Relations

In the last decade, considerable efforts have been put in annotating corpora with co-referential relations in order to support the development of co-reference resolution systems. Co-reference resolution is the task of automatically recognising which words or expressions (most often noun phrases) refer to the same discourse entity in a particular text or dialogue. The applicability of the accurate identification of co-reference relations between noun phrases is huge: in information extraction, question answering or in machine translation. Therefore, not only a widespread language such as English (e.g. ACE-2 [11], ARRAU [30], OntoNotes 3.0 [56]), but also smaller languages such as Czech (PDT 2.0; [19]) and Catalan (AnCora-Ca; [31]) can now rely on annotated resources for co-reference resolution. Through the annotation of the SoNaR-1 corpus, we created one of the largest data sets currently available to co-reference resolution research. Furthermore, the balanced nature of the data also allows for studying cross-genre performance [9].

The first Dutch corpus annotated with co-referential relations between nominal constituents was created in 2005 [15]. In the STEVIN COREA project, the annotation guidelines from [15] were refined and also extended to the labeling of bridge relations [12].[24] These COREA guidelines served as the basis for the annotation of co-reference in the SoNaR-1 corpus. The guidelines allow for the annotation of four relations and special cases are flagged. The four annotated relations are identity (NPs referring to the same discourse entity), bound, bridge (as in part-whole, superset-subset relations) and predicative. The following special cases were flagged: negations and expressions of modality, time-dependency and identity of sense (as in the so-called paycheck pronouns [18]). Co-reference links

[24] See also Chap. 7, p. 115.

were annotated between nominal constituents, which could take the form of a pronominal, named entity or common noun phrase, as exemplified in (3), (4) and (5).

3. Nederland gaat de bestrijding van het terrorisme [id="21"] anders en krachtiger aanpakken. Minister Donner van justitie krijgt verregaande bevoegdheden in die strijd [id = "2" ref="1" type="ident"].
4. Het is een eer om hier te zijn op MGIMO [id="1"]. Deze prachtige universiteit [id="2" ref="1" type="ident"] is een kweekvijver voor diplomatiek talent [id="3" ref="1" type="pred"]. Deze instelling [id="4" ref="1" type= "ident"] heeft hechte contacten met Nederland.
5. Binnen in de gymzaal [id="1"] plakken gijzelaars [id="2"] de ramen [id="3" ref="1" type="bridge"] af en plaatsen ze [id="4" ref="2" type="ident"] explosieven aan de muur [id="5" ref="1" type="bridge"].

   (English: Inside the gym, the hijackers covered the windows and attached explosives to the walls)

In order to avoid conflicts between the annotation layers, the co-reference annotations were performed on the nominal constituents, which were extracted from the manually validated syntactic dependency trees [53]. Furthermore, we checked for inconsistencies with the named entity layer. We again used MMAX2 as annotation environment.

Since inter-annotator agreement for this labeling task was already measured in the framework of the design of the annotation guidelines [12], no separate inter-annotator agreement assessment was done. Hendrickx et al. [12] computed the inter-annotator agreement on the identity relations as the F-measure of the MUC-scores [55] obtained by taking one annotation as 'gold standard' and the other as 'system output'. They report an inter-annotator agreement of 76 % F-score on the identity relations. For the bridging relations, an agreement of 33 % was reported.

Due to the low performance of the current classification-based co-reference resolution systems for Dutch [12, 15] no automatic pre-annotation was performed to support or accelerate the annotation process.

13.4.5 Annotation of Semantic Roles

The labeling of semantic roles was initiated in the D-Coi project and resulted in a set of guidelines [46] which were further extended in the SoNaR project and a small labeled data set of about 3,000 predicates. For the development of the guidelines, we considered the annotation scheme proposed within existing projects such as FrameNet [17] and PropBank [29]. Mainly because of the promising results obtained for automatic semantic role labeling using the PropBank annotation scheme, we decided to adapt the latter scheme to Dutch. In the case of traces, PropBank creates co-reference chains for empty categories while in our case, empty categories are almost non-existent and in those few cases in which they are attested, a co-indexation has been established already at the syntactic level. Furthermore, in SoNaR we assume dependency structures for the syntactic representation while PropBank employs phrase structure trees. In addition, Dutch behaves differently from English with respect to certain constructions (i.e. middle verb constructions) and these differences were also spelled out.

Besides the adaptation (and extension) of the guidelines to Dutch, a Dutch version of the PropBank frame index was created. In PropBank, frame files provide a verb specific description of all possible semantic roles and illustrate these roles by examples. The lack of example sentences makes consistent annotation difficult. Since defining a set of frame files from scratch is very time consuming, we annotated Dutch verbs with the same argument structure as their English counterparts, thus using English frame files instead of creating Dutch ones.

For the annotation of the semantic roles, we relied on the manually corrected dependency trees and TrEd[25] was used as annotation environment.

The PropBank role annotation is exemplified below, using two previously introduced examples (cf. (3) and (5)):

6. Nederland(Arg0)— gaat — de bestrijding van het terrorisme (Arg1) — anders en krachtiger (ArgM-MNR) — aanpakken (PRED). Minister Donner van justitie (Arg0)— krijgt (PRED) — verregaande bevoegdheden in die strijd (Arg1).
7. Binnen in de gymzaal (ArgM-LOC) — plakken (PRED) — gijzelaars (Arg0) — de ramen (Arg1) — af en —plaatsen (PRED)— ze (Arg0) —explosieven(Arg1)— aan de muur (Arg2).

Lacking a training corpus for Duch semantic role labeling, we initially created a rule-based tagger based on D-Coi dependency trees [24], called XARA (XML-based Automatic Role-labeler for Alpino-trees). It establishes a basic mapping between nodes in a dependency graph and PropBank roles. A rule in XARA consist of an XPath expression that addresses a node in the dependency tree, and a target label for that node, i.e. a rule is a (path, label) pair. Once sufficient training data were available, we also developed a supervised classifier, and more specifically the memory-based learning classifiers implemented in TiMBL [6], for the task. Instead of starting annotation from scratch we decided to train our classifier on the sentences annotated for D-Coi in order to pre-tag all sentences, thus rephrasing the annotation task as a verification task. After manually verifying 50,000 words we performed a first error analysis and retrained the classifier on more data in order to bootstrap the annotation process. In total, 500,000 words were manually verified. This dataset again served as the basis for the further adaptation of the classifier, which also takes into account the results of the new annotation layers of NE and co-reference. This adapted classifier labeled the remaining 500K of the SoNaR-1 corpus.

[25] URL: http://ufal.mff.cuni.cz/$\sim$pajas/tred/o

13.4.6 Annotation of Temporal and Spatial Entities

Whereas usually these two layers of annotation are handled separately, we have used STEx (which stands for Spatio Temporal Expressions), a combined spatiotemporal annotation scheme. STEx takes into account aspects of both TimeML [36] upon which the recent ISO standard ISO TimeML is mainly based[26] and SpatialML[43], serving as an ISO standard under construction. A first version of STEx, MiniSTEx, was developed within the D-Coi project, the tool used there being a semi-automatic one. Work on MiniSTEx was continued in the AMASS++-project (IWT-SBO). The resulting STEx approach is a hybrid one, which uses rules, a large spatio-temporal knowledge base, the Varro toolkit (cf. [22,23]) and TiMBL [7] to annotate texts fully automatically. The correctors are not confronted with tags with an under-treshold probability in case several tags are in se possible unless all of these are under-treshold.

Within the SoNaR project, the STEx spatial scheme was largely restricted to geospatial annotation.[27] Moreover, due to financial and temporal restrictions, we had to limit ourselves to recognition and normalisation of temporal and geospatial entities, while reasoning was ignored.

The current STEx scheme handles spatial and temporal expressions much in the same way as MiniSTEx [37–39], i.e., contrary to ISO TimeML and (ISO) SpatialML, in combination (cf. Table 13.4). We consider this quite a unique characteristic of our approach [41]. Another point in which STEx deviates from other approaches concerns the use of a feature noise. People often formulate carelessly, even journalists in quality newspapers or weeklies, for example mixing Engels (English) and Brits (British) in "de Engelse/Britse minister-president". As England is in Great Britain, would this mean that there are two prime-ministers, one of England and one of Great Britain? Or is this to be considered noisy information as in Dutch the notions England, United Kingdom and Great Britain are often mixed up? And when someone remarked the 30th of April 2011 to have been in Paris a year ago, does that mean that person was there the 30th of April 2010 (on the exact date) or rather that he or she was there around that date? In STEx such expressions come with a feature noise = "yes".

Besides the fact that STEx uses geospatial information to determine temporal information and the other way around, STEx also differs from both TimeML and SpatialML in that it is provides more details (cf. [38, 39]). In the AMASS++-project this turned out to be very useful in multidocument applications, like summarisation and information retrieval as it makes available information not expressed in a text.

8. Zij hebben hun zoon gisteren [temp type="cal" ti="tp-1" unit="day" val=
   "2008-05-22"] in Amsterdam [geo type="place" val="EU::NL::-::NH::

[26] Cf. TimeML Working Group 2010.

[27] In the ISO working group on SpatialML most attention up till now was devoted to spatial phenomena in general, not to geospatial ones.

Table 13.4 The resemblance between temporal and spatial analyses

Temporal	Geospatial
Time of perspective	Place of perspective
Time of location	Place of location
Time of eventuality	Place of eventuality
Duration	Distance
Shift of perspective	Shift of perspective
Relations	Relations

Amsterdam::Amsterdam" coord="52.37,4.9"] gezien [temp type="event" value="vtt" rel="before(ti,tp)"]
(English: They saw their son yesterday in Amsterdam)

In example (8) the time-zone associated with it (timezone = "UTF+1") is filtered out, although it is contained in the metadata coming with the text. Only when its value is overruled by a statement in the text it will be mentioned in the annotation itself. Example (8) also contains a shorthand version of the formulas we associated with several temporal expressions. ti = "tp-1" unit = "day" says that the time of eventuality ti is the time of perspective tp minus 1. As the unit involved is that of day, only that variable is to be taken into account. So, yesterday is to be associated with a formula, not with an accidental value (like "2008-05-22" in (8)). In a second step, the calculations are to be performed. This is crucial for a machine learning approach: not the value for yesterday is to be learned, but the formula associated with it.

In the context of the SoNaR corpus, STEx made use of the information available through previous syntactic and semantic layers.[28,29] In some cases it completed and disambiguated such information. For example, the location related annotations at the level of NER would be disambiguated. When a sentence like (8) occurred in a document, usually an expression like Amsterdam could be disambiguated, stating that the instantiation of Amsterdam meant was the town of Amsterdam in the Netherlands, not one of the towns or villages in the US, Canada, ... Especially in a corpus, the metadata coming with a file allow for such an annotation (cf. [38]). Co-reference was also very useful, the same holds especially for metonymy as annotated in NER (cf. also [20]). As remarked above, spatio-temporal annotation in SoNaR was performed (semi-)automatically, using a large knowledge base containing geospatial and temporal data, combinations of these and especially also cultural data with respect to such geospatial and temporal data. Cultural aspects like tradition (Jewish, Christian), geographical background, social background have their effects on the (intended) interpretation of temporal and geospatial data (cf. Fig. 13.2) by

[28] With regard to the exception of the Semantic Role Labeling (SRL) which was ignored, as for practical reasons SRL and STEx were performed in parallel.

[29] In the AMASS++ project [40] a version of STEx was used in which it had to rely on automatic PoS tagging and chunking. In a future paper we intend to compare such approaches: is manual correction/addition of further layers of annotation worth the effort (time and money)?

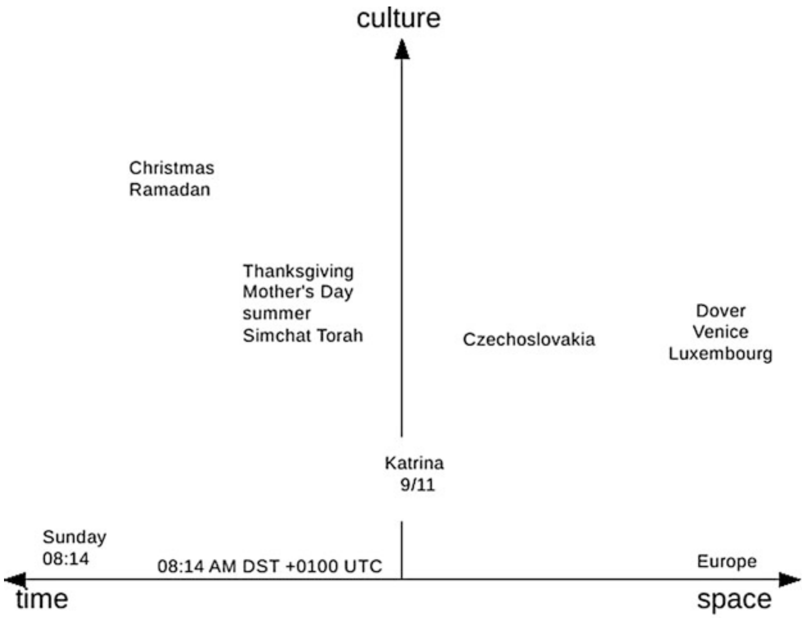

Fig. 13.2 Eventualities with temporal, geospatial or and/or cultural aspects

the people meant to read a specific text. For example: what is considered as the begin and end dates of World War II is not the same all over Europe and the rest of the world.[30] The same holds for the date(s) associated with Christmas, or Thanksgiving. Or to decide which Cambridge (UK, US) is referred to, or which Antwerpen (Antwerp): the province, the municipality or the populated place.[31] Each annotation was in principle corrected by one corrector (student), some substantial parts were corrected by more students in order to ensure annotator agreement. The time needed for correcting a file depended on the type of file, even on its topic. Legal texts for example, we found, were rather easy. However, the description of the history of a few Dutch hamlets over the last 500 years or the ins and outs of the American Civil War might take very long as in those cases the knowledge base will not contain all the relevant data.

[30] With regard to begin date: September 1939 (invasion of Poland), May 1940 (invasion of The Netherlands and Belgium), December 1941 (US, Pearl Harbor). Or . . . ?

[31] At the moment, the precision for such geospatial anchors in STEx is 0.92, recall 0.91 (small scale test for some 200 instances).

13.5 Concluding Remarks

While the Spoken Dutch Corpus already provided researchers with spoken language data, at the start of the STEVIN programme the dire need for a large resource for written data persisted. Through investment in the D-Coi and SoNaR projects directed at the construction of a 500 MW corpus an important gap in the Dutch language resources infrastructure was filled. But the impact of these projects extends well beyond the delivery of the 500 MW reference corpus as significant contributions were made to the development and consolidation of de facto standards, and tools and procedures were developed that were also used in various other projects.[32]

Although the D-Coi project was defined as a preparatory project which aimed to develop the procedures, protocols and tools needed for the construction of a large corpus, one of the more tangible results for the end-user was the 54 MW pilot corpus that was compiled [26]. In order to facilitate corpus exploitation, COREX – the corpus exploitation software developed for use with the Spoken Dutch Corpus – was adapted so that with one and the same tool both the Spoken Dutch corpus and the D-Coi corpus can now be accessed. The D-Coi corpus and the exploitation software are available through the Dutch HLT Agency.[33]

Through the SoNaR project two further corpora have become available: the SoNaR-500 corpus and the SoNaR-1 corpus. The SoNaR-500 corpus is available in two formats, the D-Coi+ format and the latest development FoLiA (Format for Linguistic Annotation; [51]). With the D-Coi+ format we are compatible with previous (intermediate) releases of the corpus. However, as the D-Coi+ format is not capable of accommodating the annotations for NE and has no provisions for specific characteristics associated with data from the new media, we have decided to adopt FoLiA for which this is not a problem. The annotations for the SoNaR-1 corpus are available in the formats as they were produced, i.e. MMAX for co-reference and named entities, TrEd for semantic roles, STEx XML for temporal and spatial entities.

For the exploitation of the 500 MW reference corpus presently no exploitation software is available, nor is the development of such software presently foreseen. For the exploitation of the SoNaR-1 corpus dedicated tools are already available for the syntactic annotation (cf. Chap. 9, p. 147), while currently in the context of

[32] Standards developed in D-Coi and SoNaR have been used in for example the STEVIN Jasmin-CGN and Dutch Parallel Corpus projects but also in the NWO-funded BasiLex and Dutch SemCor projects. As for tools and procedures, the corpus clean-up procedure developed by Reynaert has been adopted in the NWO-funded Political Mashup project and a project funded by CLARIN-NL, viz. VU-DNC, while it is also available as a web application/servide in the CLARIN infrastructure. Experiences in the D-Coi project have guided the development of by now widely used tools such as the Tilburg tagger/lemmatisers and the Alpino parser.

[33] With additional funds from NWO the HLT Agency together with Polderland Language and Speech Technology bv continued to develop the tool. The aim was to make corpora accessible over the internet and to make possible the exploitation of other corpora (such as JASMIN-CGN).

the TTNWW project all the tools and the semantic annotations discussed in this chapter will be made more easily accessible, especially for researchers in human and social sciences.[34] Apart from the D-Coi pilot corpus and the SoNaR-500 and the SoNaR-1 corpora, there are large quantities of surplus materials. As observed in Sect. 13.2.2, to the extent possible within the limitations of the SoNaR project, these data have been processed. Of the materials that presently remain in their original form a substantial part is in PDF. In our experience it is advisable to leave these data be until such a time when at some point in the future there is a breakthrough in the text extraction technology which makes it possible to extract text from PDF without losing valuable information.[35]

Acknowledgements Thanks are due to our collaborators in these projects (in random order): Paola Monachesi, Gertjan van Noord, Franciska de Jong, Roeland Ordelman, Vincent Vandeghinste, Jantine Trapman, Thijs Verschoor, Lydia Rura, Orphée De Clercq, Wilko Apperloo, Peter Beinema, Frank Van Eynde, Bart Desmet, Gert Kloosterman, Hendri Hondorp, Tanja Gaustad van Zaanen, Eric Sanders, Maaske Treurniet, Henk van den Heuvel, Arjan van Hessen, and Anne Kuijs.

Appendix

In the first column of Table A.1 the various corpus components and text types are listed. The second column indicates the data volumes foreseen in the original design. The third column shows the data volumes in the D-Coi pilot corpus. The remaining three columns give the data volumes actually realised in the SoNaR-500 corpus. NLD stands for data originating from the Netherlands, BEL for data from Flanders, and OTH for data whose origin could not be established. Data volumes are in millions of words.

[34] The acronym TTNWW stands for TST Tools voor het Nederlands als Webservices in een Workflow (HLT Tools for Dutch as Web Services in a Work Flow). This Flemish-Dutch pilot project is financed by the Flemish (Department of Economy, Science and Innovation) and Dutch (via CLARIN-NL) governments.

[35] For a recent appraisal of the state of the art in PDF text extraction technology we refer to a recent technical paper released by Mitre [13]. The main conclusion there 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.

Table A.1

			SoNaR-500		
	Original design	D-Coi	NLD	BEL	OTH
Written to be read, published, electronic	177.5	27.3	36.8	59.2	32.8
Written to be read, published, printed	185.0	25.1	101.4	233.9	19.5
Written to be read, unpublished, electronic	100.0	0	1.6	11.4	0
Written to be read, unpublished, printed	10.0	0	0	0	0
Written to be read, unpublished, typed	20.0	0	0	0	0
Written to be spoken, unpublished, electronic	2.5	0.9	2.8	25.3	0
Written to be spoken, unpublished, typed	5.0	0.7	0.7	0	0

References

1. Aston, G., Burnard, L.: The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh (1998)
2. Braasch, A., Farse, H., Jongejan, B., Navaretta, C., Olsen, S., Pedersen, B.: Evaluation and Validation of the D-Coi Pilot Corpus. Center for Sprokteknologi, Copenhagen (2008)
3. Carletta, J.C.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996)
4. Chinchor, N., Robinson, P.: MUC-7 Named Entity Task Definition (version 3.5) (1998)
5. Daelemans, W., Strik, H.: Het Nederlands in de taal-en Spraaktechnologie: prioriteiten Voor Basisvoorzieningen. Nederlandse Taalunie, The Hague (2002)
6. Daelemans, W., van den Bosch, A.: Memory-Based Language Processing. Cambridge University Press, Cambridge (2005)
7. Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A.: TiMBL: tilburg memory based learner, version 5.1.0, reference guide. Technical Report ILK 04-02, ILK Research Group, Tilburg University (2004)
8. De Clercq, O., Reynaert, M.: SoNaR acquisition manual version 1.0. Technical Report LT3 10-02, LT3 Research Group – Hogeschool Gent (2010). http://lt3.hogent.be/en/publications/
9. De Clercq, O., Hoste, V., Hendrickx, I.: Cross-domain Dutch coreference resolution. In: Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing. RANLP 2011, Hissar, Bulgaria (2011)

10. Desmet, B., Hoste, V.: Named entity recognition through classifier combination. In: Computational Linguistics in the Netherlands 2010: Selected Papers from the Twentieth CLIN Meeting, Utrecht (2010)
11. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, R., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, pp. 837–840. LREC-2004 (2004)
12. Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Vloet, J.V.D., Verschelde, J.L.: A coreference corpus and resolution system for Dutch. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech, pp. 144–149. LREC-2008 (2008)
13. Herceg, P.M., Ball, C.N.: 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 (2011). http://www.mitre.org/work/tech_papers/2011/11_0753/11_0753.pdf
14. Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I., Van der Wouden, T.: CGN syntactische annotatie. http://www.ccl.kuleuven.be/Papers/sa-man_DEF.pdf (2004)
15. Hoste, V.: Optimization issues in machine learning of coreference resolution. Ph.D. thesis, Antwerp University (2005)
16. Ide, N., Macleod, C., Fillmore, C., Jurafsky, D.: The American national corpus: an outline of the project. In: Proceedings of International Conference on Artificial and Computational Intelligence. ACIDCA-2000, Monastir (2000)
17. Johnson, C.R., Fillmore, C.J., Petruck, M.R.L., Baker, C.F., Ellsworth, M.J., Ruppenhofer, J., Wood, E.J.: FrameNet: theory and practice. ICSI Technical Report tr-02-009 (2002)
18. Karttunen, L.: Discourse Referents. Syntax and Semantics, vol. 7. Academic, New York (1976)
19. Kučova, L., Hajičova, E.: Coreferential relations in the Prague dependency treebank. In: Proceedings of DAARC 2004, Azores, pp. 97–102 (2004)
20. Leveling, J., Hartrumpf, S.: On metonymy recognition for geographic information retrieval. Int. J. Geogr. Inf. Sci. **22**(3), 289–299 (2008)
21. Markert, K., Nissim, M.: Towards a corpus annotated for metonymies: the case of location names. In: Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, pp. 1385–1392. LREC-2002 (2002)
22. Martens, S.: Varro: an algorithm and toolkit for regular structure discovery in treebanks. In: Proceedings of Coling 2010, Beijing, pp. 810–818 (2010)
23. Martens, S.: Quantifying linguistic regularity. Ph.D. thesis, KU Leuven (2011)
24. Monachesi, P., Stevens, G., Trapman, J.: Adding semantic role annotation to a corpus of written Dutch. In: Proceedings of the Linguistic Annotation Workshop (Held in Conjunction with ACL 2007), Prague (2007)
25. Oostdijk, N.: The spoken dutch corpus. Outline and first evaluation. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, pp. 887–894. LREC-2000 (2000)
26. Oostdijk, N.: Dutch language corpus initiative, pilot corpus. Corpus description. TR-D-COI-06-09 (2006)
27. Oostdijk, N.: A reference corpus of written Dutch. Corpus design. TR-D-COI-06f (2006)
28. Oostdijk, N., Boves, L.: User requirements analysis for the design of a reference corpus of written Dutch. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, pp. 1206–1211. LREC-2006 (2006)
29. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: a corpus annotated with semantic roles. Comput. Linguist. J. **31**(1) (2005)
30. Poesio, M., Artstein, R.: Anaphoric annotation in the ARRAU corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, Marrakech, pp. 1170–1174. LREC-2008 (2008)
31. Recasens, M., Marti, M.A.: AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. Lang. Resour. Eval. **44**(4), 315–345 (2010)

32. Reynaert, M.: Corpus-induced corpus cleanup. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC-2006, Trento, pp. 87–92 (2006)
33. Reynaert, M.: Non-interactive OCR post-correction for giga-scale digitization projects. In: Gelbukh, A. (ed.) Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008, vol. 4919, pp. 617–630. Springer, Berlin (2008)
34. Reynaert, M.: Character confusion versus focus word-based correction of spelling and OCR variants in corpora. Int. J. Doc. Anal. Recognit. 1–15 (2010). http://dx.doi.org/10.1007/s10032-010-0133-5, doi:10.1007/s10032-010-0133-5
35. Sanders, E.: Collecting and analysing chats and tweets in SoNaR. In: Proceedings of the Eighth International Conference of Language Resources and Evaluation, Istanbul, pp. 2253–2256. LREC-2012 (2012)
36. Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML annotation guidelines, version 1.2.1. http://timeml.org/site/publications/specs.html (2006)
37. Schuurman, I.: Spatiotemporal annotation on top of an existing treebank. In: De Smedt, K., Hajic, J., Kuebler, S. (eds.) Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, Bergen, pp. 151–162 (2007)
38. Schuurman, I.: Which New York, which Monday? The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions. In: Proceedings of CLIN 17, Leuven (2007)
39. Schuurman, I.: Spatiotemporal annotation using MiniSTEx: how to deal with alternative, foreign, vague and obsolete names? In: Proceedings of the Sixth Conference on International Language Resources and Evaluation (LREC'08), Marrakech (2008)
40. Schuurman, I., Vandeghinste, V.: Cultural aspects of spatiotemporal analysis in multilingual applications. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta (2010)
41. Schuurman, I., Vandeghinste, V.: Spatiotemporal annotation: interaction between standards and other formats. In: IEEE-ICSC Workshop on Semantic Annotation for Computational Linguistic Resources, Palo Alto (2011)
42. Schuurman, I., Schouppe, M., Van der Wouden, T., Hoekstra, H.: CGN, an annotated corpus of spoken Dutch. In: Proceedings of the Fourth International Conference on Linguistically Interpreed Corpora, Budapest, pp. 101–112. LINC-2003 (2003)
43. SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language. MITRE (2007). Version 2.0, LDC, Upenn
44. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, pp. 2142–2147. LREC-2006 (2006) http://arxiv.org/ftp/cs/papers/0609/0609058.pdf
45. Tjong Kim Sang, E.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of the 6th Conference on Natural Language Learning, Taipei, pp. 155–158 (2002)
46. Trapman, J., Monachesi, P.: Manual for semantic annotation in D-Coi. Technical Report, Utrecht University (2006)
47. Treurniet, M., De Clercq, O., Oostdijk, N., Van den Heuvel, H.: Collecting a corpus of Dutch SMS. In: Proceedings of the Eighth International Conference of Language Resources and Evaluation, Istanbul, pp. 2268–2273. LREC-2012 (2012)
48. Van den Bosch, A., Schuurman, I., Vandeghinste, V.: Transferring PoS-tagging and lemmatisation tools from spoken to written Dutch corpus development. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa. LREC-2006 (2006)
49. Van den Bosch, A., Busser, B., Canisius, S., Daelemans, W.: An efficient memory-based morphosyntactic tagger and parser for Dutch. In: Dirix, P., Schuurman, I., Vandeghinste, V., Van Eynde, F. (eds.) Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting, Leuven, pp. 99–114 (2007)

50. Van Eynde, F.: Part of speech tagging en lemmatisering. Protocol voor annotatoren in D-Coi. Centrum voor Computerlinguïstiek, Leuven. http://www.let.rug.nl/vannoord/Lassy/POS-manual.pdf internal document
51. van Gompel, M.: Folia: format for linguistic annotation. http://ilk.uvt.nl/folia/folia.pdf (2011)
52. Van Noord, G.: At last parsing is now operational. In: Verbum Ex Machina, Actes De La 13e Conference sur Le Traitement Automatique des Langues Naturelles, Leuven, pp. 20–42. TALN-2006 (2006)
53. Van Noord, G., Schuurman, I., Vandeghinste, V.: Syntactic annotation of large corpora in STEVIN. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, pp. 1811–1814. LREC-2006 (2006)
54. Van Rijsbergen, C.: Information Retrieval. Buttersworth, London (1979)
55. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, pp. 45–52 (1995)
56. Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., Houston, A.: OntoNotes Release 3.0. LDC2009T24. Linguistic Data Consortium (2009)
57. Woordenlijst Nederlandse Taal: SDU Uitgevers, The Hague (1995)