SoNaR User Documentation

version 1.0.4

Nelleke Oostdijk
Martin Reynaert
Véronique Hoste
Henk van den Heuvel

September 20, 2013

Contents

0.1 Preface

The present document describes the results available from the STEVIN-funded SoNaR project (2008-2011). The project aimed at the construction of a 500-million-word reference corpus of contemporary written Dutch for use in different types of linguistic (incl. lexicographic) and HLT research and the development of applications. The project built on the results obtained in the D-Coi and Corea projects which were awarded funding in the first call of proposals within the Stevin programme.

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. [5] found that Dutch, compared to other languages, was lagging behind. While the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; [18]) 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 D-Coi and the SoNaR project were funded by the NTU/STEVIN programme (http://www.taalunieversum.org/ stevin) under grant numbers STE04008 and
STE07014 respectively. 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, Orphe 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.

The main paper describing SoNaR is chapter 13 in the Open Access book about the Stevin programme projects [17]. A paper looking beyond SoNaR was presented at LREC-2012 [22].

The SoNaR corpus will be distributed by the Dutch-Flemish HLT Agency (TST-Centrale).

Contact:

Dr Nelleke Oostdijk
Radboud University Nijmegen
Erasmusplein 1
NL-6525 HT Nijmegen
The Netherlands

N.Oostdijk@let.ru.nl

Chapter 1

Introduction

1.1 Introduction to SoNaR

The STEVIN SoNaR project has resulted in two datasets, viz. SoNaR-500 and SoNaR-1.

SONAR-500 contains over 500 million words (i.e. word tokens) of full texts from a wide variety of text types including both texts from conventional media and texts from the new media. All texts except for texts from the social media (Twitter, Chat, SMS) have been tokenized, tagged for part of speech and lemmatized, while in the same set the Named Entities have been labelled. In the case of SoNaR-500 all annotations were produced automatically, no manual verification took place.

SoNaR-1 is a dataset comprising one million words. Although largely a subset of SoNaR-500, SoNaR-1 includes far fewer text types. With SoNaR-1 different types of semantic annotation have been provided, viz. named entity labelling, annotation of co-reference relations, semantic role labelling and annotation of spatial and temporal relations. All annotations have been manually verified.

The SoNaR project was carried out by Katholieke Universiteit Leuven (CCL), Hogeschool Gent (Dept. Vertaalkunde, LT3), Radboud University Nijmegen (CLST), Tilburg University (TiCC/ILK), Twente University (HMI), and Utrecht University (UiL-OTS). It was coordinated by Radboud University.

1.2 SoNaR: Introductory Overview

The SoNaR corpus is a corpus of contemporary standard written Dutch as encountered in texts (i.e. stretches of running discourse) originating from

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

Table 1.1: 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)

the Dutch speaking language area in Flanders and the Netherlands as well as Dutch translations published in and targeted at this area. The corpus was designed to comprise 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.

The sheer size of the corpus has made it possible to include 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 Table 1.1 a global overview is given of the composition of the corpus in terms of the three main design criteria, viz.

1. the intended delivery of the texts included,

2. whether they were published or not, and

3. their primary mode.

Table 1.2 gives an overview of the distribution of the data by country of origin per corpus component.

Corpus Component	NLD	BEL	OTH	Total
WR-P-E	36,846,020	74,578,516	32,784,094	144,208,630
WR-P-P	101,444,035	233,894,017	19,458,388	354,798,440
WR-U-E	1,563,265	11,391,742	0	26,509,835
WS-U-E	2,830,877	25,268,159	0	28,099,036
WS-U-T	676,062 0	0	0	676,062
total	143,362,259	345,132,434	52,242,482	554,292,003

Table 1.2: Corpus composition in number of word tokens split according to regional origin (NLD=originating from the Netherlands; BEL=originating from Flanders, OTH=from uncertain origin)

In the course of the SoNaR project the corpus design originally conceived was modified. 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 large quantities of this type of material. In other cases, such as 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). 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.

For each text in the corpus, particular information about the text category and license type is provided in the metadata file which accompanies the text file. More information can be found in Chapter 6.

In the next chapters, both SONAR-500 and SONAR-1 are described in more detail.

Chapter 2

SonaR: Corpus Composition

2.1 SoNaR-500

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 with 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) 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.

The contents of the corpus as well as the nature of the annotations provided were to largely determined by the needs of ongoing and projected research and development in the fields of corpus-based natural language processing. Applications such as information extraction, question-answering, document classification, and automatic abstracting that are based on underlying corpus-based techniques are expected to benefit from the large-scale analysis of particular features in the corpus. Apart from supporting corpus-based modeling, the corpus constitutes a test bed for evaluating applications, whether or not these applications are corpus-based.

2.2 SoNaR-1

2.2.1 Introduction

In order to know the exact composition of SoNaR-1 we must have a closer look at the origins of the text material included. All text material comes from the Lassy Klein corpus, but this corpus has, in turn, inherited data from previous

Corpus	#Texts	#Word tokens
DPC	167	195,791
D-Coi	587	716,649
Wikipedia	107	87,997
SONAR-1	861	1,000,437

Table 2.1: Composition of the SoNaR-1 corpus. In all SoNaR-1 comprises 1,000,437 words

projects. The texts included come from two other STEVIN-funded projects, i.e. DPC and D-Coi, and from Dutch Wikipedia. Within the Dutch Parallel Corpus (DPC) [1] project a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French was built. DPC has its proper one-million-word subset in which all processing layers have been manually verified. The DPC-data included in SoNaR-1 comes from this perfect subset. Most texts in SoNaR-1 come from the D-Coi project which was actually a pilot project for SoNaR. Finally, within Lassy Klein texts were included from Dutch Wikipedia which are now also a part of SoNaR-1.

In Table 2.1, the total number of texts and tokens included in SoNaR 1 are presented.

Following the subdivision of the global reference corpus, the texts included in SoNaR-1 can also be divided in various text types. In Table 2.2 one can find an overview of this genre subdivision.

2.2.2 Corpus annotation

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.

[1] http://www.kuleuven-kulak.be/DPC

Text type	SoNaR Code	Origin	#Word tokens
Autocues	WS-U-E-A	D-Coi	205,040
Books	WR-P-P-B	D-Coi	2,008
Brochures	WR-P-P-C	D-Coi, DPC	88,451
E-magazines, E-Newsletters	WR-P-E-C, WR-P-E-E	D-Coi	12,769
Guides, Manuals	WR-P-P-E	D-Coi, DPC	28,410
Legal texts	WR-P-P-G	D-Coi	6,468
Magazines	WR-P-P-I	D-Coi, DPC	142,840
Minutes	WR-U-T-A	DPC	1,655
Newsletters	WR-P-P-E	D-Coi, DPC	8,543
Newspapers	WR-P-P-H	D-Coi, DPC	81,130
Policy documents	WR-P-P-J	D-Coi	30,021
Press releases	WR-P-E-J	D-Coi, DPC	22,261
Proceedings	WR-P-P-K	D-Coi, DPC	14,396
Reports	WR-P-P-L	D-Coi, DPC	30,751
Speeches	WS-U-E-B	DPC	17,320
Websites	WR-P-E-H	D-Coi, DPC	47,841
Wikipedia	WR-P-E-J	D-Coi, Wikipedia	260,533
Grand total			1,000,437

Table 2.2: Composition of the SoNaR-1 corpus. In all SoNaR-1 comprises 1,000,437 words

Annotation of named entities

The (manual) annotation of a one-million-word subset of the corpus was undertaken so as 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 recognizers. The labeled data set substantially differs from the CoNLL-2002 shared task [31] data set. 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, organization and miscellaneous), a finer granularity of the named entities was aimed for and a distinction was made between the literal and metonymic use of the entities. For the development of the guidelines, the annotation schemes developed in the ACE [9] and MUC [4] programmes were taken into account, and the work on metonymy by [15]. In the resulting annotation guidelines, the focus was 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.

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 organize the fight against terrorism in a different and more powerful way. Minister of Justice Donner was given far-reaching powers in that battle.)

(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 annotation environment.

Annotation of co-reference relations

The first Dutch corpus annotated with co-referential relations between nominal constituents was created in 2005 [12]. In the STEVIN COREA project, the annotation guidelines from Hoste (2005) were refined and also extended to the labeling of bridge relations [11].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, cf. [14]. Co-reference links were annotated between nominal constituents, which could take the form of a pronominal, named entity or common noun phrase, as exemplified below.

(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 [34]. Also, the annotations were checked for inconsistencies with the named entity layer. MMAX2 was used as annotation environment.

Annotation of semantic roles

For the annotation of semantic roles the PropBank annotation scheme [20] was adapted. There are two important differences between the original PropBank scheme and the adapted annotation scheme used to annotated SoNaR-1. First, in the case of traces, PropBank creates co-reference chains

for empty categories while in the SoNaR scheme, 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. Second, in SoNaR dependency structures for the syntactic representation are assumed while PropBank employs phrase structure trees. Also worth noting is that Dutch behaves differently from English with respect to certain constructions (i.e. middle verb constructions). Therefore, these differences have also been spelled out.

Examples:

(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).

For the annotation of the semantic roles, we relied on the manually corrected dependency trees. TrEd[2] was used as annotation environment.

In total, 500,000 words were manually verified. On this data the classifier was based, which also takes into account the results of the new annotation layers of NE and co-reference. This classifier labeled the remaining 500K of the SoNaR-1 corpus.

Annotation of temporal and spatial entities

For the annotation of spatial and temporal entities [25], a combined spatiotemporal annotation scheme was used: STEx (which stands for Spatio Temporal Expressions). STEx [26] takes into account aspects of both TimeML [23] upon which the recent ISO standard ISO TimeML[3] is mainly based and SpatialML [30], serving as an ISO standard under construction.

The annotation 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 by the people meant to read a specific text. For example: what is considered as the begin and end dates of World War II

[2]http://ufal.mff.cuni.cz/tred

[3]Cf. TimeML Working Group 2010

is not the same all over Europe and the rest of the world.[4] 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.

Each annotation was in principle corrected by one corrector (student), some substantial parts were corrected by more students in order to ensure annotator agreement.

Example:

(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::Amsterdam::Amsterdam" coord="52.37,4.9"] gezien [temp type="event" value="vtt" rel="before(ti,tp)"]
(English: They've seen 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.

[4]September 1939 (invasion of Poland), May 1940 (invasion of The Netherlands and Belgium), December 1941 (US, Pearl Harbor). Or ?

Chapter 3

Organization of the SoNaR Corpus distribution

3.1 A note on file names

Each of the texts included in the corpus is given a unique ID which is used as a file name. The ID can be decomposed into five parts which provide information relating to the corpus design (cf. Section 2), viz.

1. whether a text can be classified as intended as written to be read (WR) or written to be spoken (WS);
2. whether the text was published (P) or unpublished (U);
3. whether the text appeared in electronic form (E), was printed (P), or typed (T);
4. the text type (e.g. discussion list, books, reports, etc.) using a single letter; for an overview of the different text types, see the Appendix;
5. a unique number consisting of ten digits.

3.2 SoNaR Directory structure

3.2.1 Introduction

The directory structure of the SoNaR Corpus can be represented in the form of the following tree:

In the root directory there are the following subdirectories:

- DOC/
- SONAR500/
- SONAR1/

This structure is explained in the next subsections.

3.2.2 Directory DOC/

The DOC directory contains the general end-user documentation about SONAR-500 and SONAR-1

- This file: SoNaR_end-user_documentation_Final.pdf
- Part of Speech tagging and lemmatization of the D-Coi Corpus: D-COI-05-01-POS_manual_VanEynde.pdf
- Technical specification of File formats and validation tool in D-Coi: D-COI-06-02.pdf

3.2.3 Directory SONAR500/

SONAR500 has subdirectories:

- DATA/
- LISTS/

LISTS/ contains 11 further subdirectories. The first 10 directories contain the n-gram frequency lists for the SoNaR-500 corpus:

- 1gms/ : single words
- 1gmstotal/ : single word frequencies over the whole corpus
- 2gms/ : combinations of 2 adjacent words
- 2gmstotal/ : bigram frequencies over the whole corpus
- 3gms/ : combinations of 3 adjacent words
- 3gmstotal/ : trigram frequencies over the whole corpus
- 4gms/ : combinations of 4 adjacent words

- 4gmstotal/ : 4-gram frequencies over the whole corpus
- 5gms/ : combinations of 5 adjacent words
- 5gmstotal/ : 5-gram frequencies over the whole corpus. These lists are hapaxed: 5-grams occurring only once have not been retained due to memory restrictions
- contents/: contains overview lists detailing country provenance (Belgium or the Netherlands) per text category

These frequency lists have been produced for each separate text category available in SoNaR:

- wordfreqlist : contains the word type frequencies
- lemmafreqlist: contains the accumulated lemma frequencies
- lemmaposfreqlist: contains the accumulated frequencies for the combinations of lemmas and pos-tags.

The file names are composed as follows:

- < textcat >.wordfreqlist.ngms.tsv
- < textcat >.lemmafreqlist.ngms.tsv
- < textcat >.lemmaposfreqlist.ngms.tsv

'tsv' stands for comma separated values: the columns are separated by TABs.

There are also frequency lists covering the full SONAR-500 corpus. These are named as follows, where 'x' denotes the n-gram:

- SONAR500.wordfreqlist.x-gram.total.tsv
- SONAR500.lemmafreqlist.x-gram.total.tsv
- SONAR500.lemmaposfreqlist.x-gram.total.tsv

The last subdirectory contents/ contains lists of files from the various language regions to assist the user selecting these without reference to the CMDI metadata files in which this information is naturally also available on a per text basis. For each text category there are three extra files: < textcat >-BEL,NLD,OTH.txt. Each of these files is a list of the texts in the directory that are from Dutch (NLD), Belgian (BEL) or OTHer language origin, respectively. The labels NLD and BEL are taken from the country codes ISO-3166-3. They reflect the geographical origin of a text (rather than the language origin which is typically Dutch in both cases). OTH includes the texts with unknown data origin or containing a combination of languages in one text, such as SMS or tweets may have. The list contains all the file names from the directory that are from the specific language region for the DCOI, FoLiA and CMDI files.

The files are meant to allow users to quickly sort and select the data according to language region in the absence of the means to parse the CMDI metadata files.

DATA/ is a directory with the following structure:

Text category: WR-P-E-A_discussion_lists, ..WS-U-T-B_texts_for_the_visually_impaired

The scheme of the directory structure for SONAR-500 is thus:

- SONAR500/
  - LISTS/
  - DATA/
    - WR-P-E-A_discussion_lists/
    - ...
    - WS-U-T-B_texts_for_the_visually_impaired/

We give a full overview of the directory names for all text categories as well as overviews of numbers of files and the space they take in in compressed and expanded forms in Section 3.3.

The leaf directories contain the data files in FoLiA format, in DCOI+ format and the corresponding metadata files in CMDI format. Thus there are three files per text. DCOI formatted files, however, are absent for the new media text categories: WR-E-P-L_tweets, WR-U-E-A_chats, WR-U-E-D_SMS.

The names of the data files in the leaf part of the tree are composed as follows: < text category >-< 10digitnumber >.< file type: folia, dcoi, cmdi >.xml

Examples:

WR-P-E-J-0000000014.folia.xml
WR-P-E-J-0000000014.dcoi.xml
WR-P-E-J-0000000014.cmdi.xml

Each directory at text category level with FoLiA files also contains an XSLT scheme file (sonar-foliaviewer.xsl). This file allows the user to open a FoLia XML file in a browser with a user-friendly view (without XML-codes) of the text contained in the file.

3.2.4 Directory SONAR1/

For the SONAR1 subcorpus the direct subdirectories are:

- DOC/
- COREF/
- NE/
- SRL/
- SPT/
- POS/

DOC/ contains files with general information about the composition of the corpus and a list of files.

COREF/ (=Coreference annotations) has the subdirectories:

- Documentation/
- SONAR_1_COREF/
  - Documentation/ contains the annotation manuals and papers about the annotation.
  - SONAR_1_COREF/ contains the MMAX files and subdirectories needed to view the files using the MMAX2 tool.

NE/ (=Named Entities) has the subdirectories:

- Documentation/
- NERD/
- SONAR_1_NE/
  - Documentation/ contains the annotation manuals and papers about the annotation.
  - NERD/ is the directory with the tool used for Named Entity recognition.
  - SONAR_1_NE/ contains the subdirectories:
    * IOB/
    * MMAX/

    These directories contain the annotated text in the MMAX and IOB formats respectively.

SRL/ (= Semantic Roles) has the subdirectories:

- Documentation/
- SONAR_1_SRL/
  - Documentation contains the annotation manuals and papers about the annotation. It further contains the Alpino_for_TrEd/ directory with special provisions for TrEd to read in the Alpino xml format.
  - SONAR_1_SRL/ holds the subdirectories:
    * AUTO500/
    * MANUAL500/
      · AUTO500/ contains the 500K corpus that has been automatically labeled.
      · MANUAL500/ contains the 500K corpus that has been completely manually verified.

SPT/ (=Spatiotemporal annotations) has the subdirectories:

- Documentation/
- SONAR_1_STEx/
  - Documentation contains the annotation manual and papers about the annotation. XML files are in Alpino XML with occasional extra features and tags for spatiotemporal annotations.

POS/ : contains the handcrafted POS tags of the SONAR1 part that were created in the D-Coi project.

3.3 SoNaR-500: Overview of file names, file sizes and numbers

The contents of the archive files in the SoNaR distribution containing the text files in FoLiA format and their metadata in CMDI xml format both unpack to the same subdirectory per text type in the parent directory SONAR500/DATA/.

Be advised that these are large collections of files and assure yourself that the file system on your computer can effectively handle these amounts of files in single directories.

Further be advised that in order to dearchive these collections, you will require sufficient amounts of storage space on your computer hard disk. The space required to unpack all text files is almost 500 gigabytes.

We next provide overviews of the space requirements involved for both FoLiA and CMDI xml files per text type.

Table 3.1 gives an overview of the sizes of the FoLiA xml directories per text type, as well as of the numbers of text files contained within these directories.

Table 3.2 gives an overview of the sizes of the metadata files for the SoNaR corpus in the CMDI xml directories per text type, as well as of the numbers of text files contained within these directories.

Table 3.3 gives an overview of the sizes of the frequency files for the SoNaR corpus in the directories per n-gram type, as well as of the numbers of frequency files contained within these directories.

Archive files	Size in bytes	Human-readable size	De-archives in directory	# FoLiA files	De-archived size in bytes	De-archived human-readable size
SoNaR500.Curated.WR-P-E-A_discussion_lists.20130312.tar.gz	2662064047	2.5G	SONAR500/DATA/WR-P-E-A_discussion_lists	702091	57095516	55G
SoNaR500.Curated.WR-P-E-C_e-magazines.20130312.tar.gz	373719285	357M	SONAR500/DATA/WR-P-E-C_e-magazines	18699	8346020	8.0G
SoNaR500.Curated.WR-P-E-E_newsletters.20130312.tar.gz	91008	89K	SONAR500/DATA/WR-P-E-E_newsletters	3	1900	1.9M
SoNaR500.Curated.WR-P-E-F_press_releases.20130312.tar.gz	14822657	15M	SONAR500/DATA/WR-P-E-F_press_releases	1053	335804	328M
SoNaR500.Curated.WR-P-E-G_subtitles.20130312.tar.gz	1113851387	1.1G	SONAR500/DATA/WR-P-E-G_subtitles	8368	27156856	26G
SoNaR500.Curated.WR-P-E-H_teletext_pages.20130312.tar.gz	18831627	18M	SONAR500/DATA/WR-P-E-H_teletext_pages	93	441788	432M
SoNaR500.Curated.WR-P-E-I_web_sites.20130312.tar.gz	129267839	124M	SONAR500/DATA/WR-P-E-I_web_sites	956	3073452	3.0G
SoNaR500.Curated.WR-P-E-J_wikipedia.20130312.tar.gz	1048106005	1000M	SONAR500/DATA/WR-P-E-J_wikipedia	124124	22982840	22G
SoNaR500.Curated.WR-P-E-K_blogs.20130312.tar.gz	6271138	6.0M	SONAR500/DATA/WR-P-E-K_blogs	778	137088	134M
SoNaR500.Curated.WR-P-E-L_tweets.20130312.tar.gz	755525504	721M	SONAR500/DATA/WR-P-E-L_tweets	602	12061560	12G
SoNaR500.Curated.WR-P-P-B_books.20130312.tar.gz	1077368999	1.1G	SONAR500/DATA/WR-P-P-B_books	507	25703756	25G
SoNaR500.Curated.WR-P-P-C_brochures.20130312.tar.gz	50946113	49M	SONAR500/DATA/WR-P-P-C_brochures	86	1211328	1.2G
SoNaR500.Curated.WR-P-P-D_newsletters.20130312.tar.gz	1430112	1.4M	SONAR500/DATA/WR-P-P-D_newsletters	6	33084	33M
SoNaR500.Curated.WR-P-P-E_guides_manuals.20130312.tar.gz	9791981	9.4M	SONAR500/DATA/WR-P-P-E_guides_manuals	12	235700	231M
SoNaR500.Curated.WR-P-P-F_legal_texts.20130312.tar.gz	423405624	404M	SONAR500/DATA/WR-P-P-F_legal_texts	7860	10429908	10G
SoNaR500.Curated.WR-P-P-G_newspapers.20130312.tar.gz	9250918054	8.7G	SONAR500/DATA/WR-P-P-G_newspapers	708600	205862324	197G
SoNaR500.Curated.WR-P-P-H_periodicals_magazines.20130312.tar.gz	4022948293	3.8G	SONAR500/DATA/WR-P-P-H_periodicals_magazines	176043	90584648	87G
SoNaR500.Curated.WR-P-P-I_policy_documents.20130312.tar.gz	385014148	368M	SONAR500/DATA/WR-P-P-I_policy_documents	216	9137648	8.8G
SoNaR500.Curated.WR-P-P-J_proceedings.20130312.tar.gz	13171077	13M	SONAR500/DATA/WR-P-P-J_proceedings	18	311620	305M
SoNaR500.Curated.WR-P-P-K_reports.20130312.tar.gz	93209057	89M	SONAR500/DATA/WR-P-P-K_reports	81	2151816	2.1G
SoNaR500.Curated.WR-U-E-A_chats.20130312.tar.gz	409263732	391M	SONAR500/DATA/WR-U-E-A_chats	1304	6758044	6.5G
SoNaR500.Curated.WR-U-E-D_sms.20130312.tar.gz	22132962	22M	SONAR500/DATA/WR-U-E-D_sms	218	373324	365M
SoNaR500.Curated.WR-U-E-E_written_assignments.20130312.tar.gz	13943170	14M	SONAR500/DATA/WR-U-E-E_written_assignments	188	344640	337M
SoNaR500.Curated.WS-U-E-A_auto_cues.20130312.tar.gz	1323132	1.3G	SONAR500/DATA/WS-U-E-A_auto_cues	313158	28365208	28G
SoNaR500.Curated.WS-U-T-B_texts_for_the_visually_impaired.20130312.tar.gz	28316	28M	SONAR500/DATA/WS-U-T-B_texts_for_the_visually_impaired	944	666464	651M
Totals	22734440	22G	SONAR500/DATA/	2066008	513802340	491G

Table 3.1: Directory names, sizes and contents (in numbers of files) of the SoNaR-500 corpus in FoLiA xml format

Archive file	Size in bytes	Human-readable size	De-archives in directories	# CMDI files	De-archived size in bytes	De-archived human-readable size
SONAR500.UserVersionCMDI.AllDirs.20130710.tar.gz	286355159	274M	WR-P-E-A_discussion_lists	702093	3573452	3.5G
			WR-P-E-C_e-magazines	18699	75808	75M
			WR-P-E-E_newsletters	26	108	108K
			WR-P-E-F_press_releases	1054	4284	4.2M
			WR-P-E-G_subtitles	8368	33936	34M
			WR-P-E-H_teletext_pages	101	408	408K
			WR-P-E-I_web_sites	974	6444	6.3M
			WR-P-E-J_wikipedia	124160	503000	492M
			WR-P-E-K_blogs	778	3148	3.1M
			WR-P-E-L_tweets	602	2444	2.4M
			WR-P-P-B_books	508	2848	2.8M
			WR-P-P-C_brochures	88	360	360K
			WR-P-P-D_newsletters	6	28	28K
			WR-P-P-E_guides_manuals	12	52	52K
			WR-P-P-F_legal_texts	7860	31868	32M
			WR-P-P-G_newspapers	709417	5636324	5.4G
			WR-P-P-H_periodicals_magazines	176049	863260	844M
			WR-P-P-I_policy_documents	222	904	904K
			WR-P-P-J_proceedings	19	80	80K
			WR-P-P-K_reports	81	624	624K
			WR-U-E-A_chats	1392	124180	122M
			WR-U-E-D_sms	218	888	888K
			WR-U-E-E_written_assignments	188	764	764K
			WS-U-E-A_auto_cues	313163	2438888	2.4G
			WS-U-T-B_texts_for_the_visually_impaired	944	3832	3.8M
Totals:	286355159	274M		2067022 [a]	13307936	13G

Table 3.2: Directory names, sizes and contents (in numbers of files) of the metadata for the SoNaR-500 corpus in CMDI xml format

[a]There are in fact 1,014 more CMDI files than texts in FoLiA format. This discrepancy is explained by the fact that in between each further processing and annotation step during corpus building, the resulting XML files were validated against the SoNaR XML schema. The actual processing steps that were passed are listed in each CMDI file. Files that failed to pass one of these steps did not make it into FoLiA format.

Archive file	Size in bytes	Human-readable size	De-archives in directories	# files	De-archived size in bytes	De-archived human-readable size
SONAR500.LISTS.1gms.MRE.tar.gz	238610840	228M	SONAR500/LISTS/1gms/	75	1032796	1009 M
SONAR500.LISTS.1gmstotal.MRE.tar.gz	58582059	56M	SONAR500/LISTS/1gmstotal/	3	248512	243M
SONAR500.LISTS.2gms.MRE.tar.gz	1991769068	1.9G	SONAR500/LISTS/2gms/	75	10262404	9.8G
SONAR500.LISTS.2gmstotal.MRE.tar.gz	401727645	384M	SONAR500/LISTS/2gmstotal/	3	2013540	2.0G
SONAR500.LISTS.3gms.MRE.tar.gz	5988492495	5.6G	SONAR500/LISTS/3gms/	75	32949800	32G
SONAR500.LISTS.3gmstotal.MRE.tar.gz	968657895	924M	SONAR500/LISTS/3gmstotal/	3	5243676	5.1G
SONAR500.LISTS.4gms.MRE.tar.gz	10350924642	9.7G	SONAR500/LISTS/4gms/	75	56852224	55 GB
SONAR500.LISTS.4gmstotal.MRE.tar.gz	1291286431	1.3G	SONAR500/LISTS/4gmstotal/	3	7056924	6.8G
SONAR500.LISTS.5gms.MRE.tar.gz	13614779195	13G	SONAR500/LISTS/5gms/	75	71946124	69 GB
SONAR500.LISTS.5gmstotal.MRE.tar.gz	1339075440	1.3G	SONAR500/LISTS/5gmstotal/	3	6983328	6.7G (hapaxed)
Totals:	35394492	34G		390	194589328	177.7G

Table 3.3: Directory names, sizes and contents (in numbers of files) of the word type, lemma and lemma and POS frequency files for the SoNaR-500 corpus in tab-separated values format. Due to memory restrictions the totaled 5-gram lists derived from all the text categories have been hapaxed: 5-grams occurring only once in the whole corpus are not listed

Chapter 4

SoNaR File Formats

4.1 SoNaR-500 File formats

All texts are available in FoLiA XML format [10]. Most of the texts are also available in the legacy DCOI(+) format. The texts from the social media (chat, twitter, SMS) are solely available in FoLiA format, since this format allows a proper representation of events and time stamps in the data which is not possible in DCOI+.

DCOI+ formatted files are therefore not available for:

- WR-P-E-L_tweets
- WR-U-E-A_chats
- WR-U-E-D_SMS

Moreover, the DCOI+ format does not allow the inclusion of Named Entity annotations whereas for FoLiA this is not a problem. Named Entity annotations are therefore only available in the files in FoLiA format. Below a description is given of both DCOI+ and FoLiA formats.

As of beginning 2013, we regard the DCOI+ format as deprecated. The SoNaR-500 corpus has undergone intensive curation on the basis of the external corpus evaluation report produced by Center for Sprokteknologi, Copenhagen, Denmark. This curation has been effected on the FoLiA format. Over 2 million word tokens have been normalized one way or another: there is now a serious discrepancy between the token content of both versions. Also, the FoLiA version has undergone an extra, new and consistent annotation of lemmas and pos-tags. A further annotation layer, i.e. morphological annotation, has furthermore been added.

4.1.1 D-Coi+ format (deprecated)

The D-Coi format is an XML format conceived for SoNaRs predecessor corpus D-Coi (Dutch Language Corpus Initiative), and is the format in which SoNaRs pre-releases have been delivered. It is included as a backward-compatibility format in SoNaR-500, alongside the newer FoLiA format. The D-Coi+ format is identical to the D-Coi format except that the IMDI header information was removed. The validation of the files was done at a stage in which the files still contained the IMDI header. Metadata information is contained in CMDI files (see Section 6.2)

The D-Coi format focuses on annotation of document structure: divisions, paragraphs, sentences and word tokens and uses XML elements in an inline fashion to describe the structure of a document, benefiting from XMLs inherent hierarchical nature. The original aim of the developers was to offer a format close to the TEI format and copy a minimal subset of TEI elements [1].

The overall structure of a D-Coi document is as follows:

```
<DCOI>
        <text>
                <gap>
                        [unannotated front matter]
                </gap>
                <body>
                        [body of the text to be annotated]
                </body>
                <gap>
                        [unannotated back matter]
                </gap>
        </text>
</DCOI>
```

D-Coi documents traditionally incorporate a metadata block in the IMDI format, prior to the text element, but in SoNaR-500 it has been decided to adopt CMDI instead and to use external files separated from the text and its annotations. Therefore the D-Coi+-files in SONAR-500 do not contain this IMDI header. The body of the text would contain structural elements such as divisions (div), headers (head) paragraphs (p), sentences (s), word tokens (w), and others. An example is provided below. A notable feature of the D-Coi format is that each element is assigned a globally unique XML Identifier, which can be used in references from external sources.

```
<body>
        <div0 xml:id="WR-P-E-J-0000000001.div0.1">
          <head xml:id="WR-P-E-J-0000000001.head.1">
            <s xml:id="WR-P-E-J-0000000001.head.1.s.1">
              <w xml:id="WR-P-E-J-0000000001.head.1.s.1.w
                  .1" pos="N(soort,ev,basis,onz,stan)"
                  lemma="stemma">Stemma</w>
            </s>
          </head>
          <p xml:id="WR-P-E-J-0000000001.p.1">
            <s xml:id="WR-P-E-J-0000000001.p.1.s.1">
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.1"
                  pos="N(eigen,ev,basis,zijd,stan)" lemma
                  ="Stemma">Stemma</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.2"
                  pos="WW(pv,tgw,ev)" lemma="zijn">is</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.3"
                  pos="LID(onbep,stan,agr)" lemma="een">
                  een</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.4"
                  pos="ADJ(prenom,basis,zonder)" lemma="
                  ander">ander</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.5"
                  pos="N(soort,ev,basis,onz,stan)" lemma=
                  "woord">woord</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.6"
                  pos="VZ(init)" lemma="voor">voor</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.7"
                  pos="N(soort,ev,basis,zijd,stan)" lemma
                  ="stamboom">stamboom</w>
              <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.8"
                  pos="LET()" lemma=".">.</w>
            </s>
                  </p>
```

The D-Coi+ format includes the ability for linguistic annotation by means of part of speech tags and lemmas using the XML attributes pos and lemma respectively. This is shown in the above example.

SoNaR documents in the D-Coi+ format are all encoded in iso-8859-15 character encoding.

The documentation of the original D-Coi format can be found at `http://lands.let.ru.nl/projects/d-coi/Doc/Voorstel_XML_basisformaat.doc` See also [1].

4.1.2 FoLiA

SoNaR-500 is delivered in FoLiA XML format `http://ilk.uvt.nl/folia` [10]. SoNaR documents in FoLiA use the unicode (UTF-8) standard.

FoLiA is an XML-based Format for Linguistic Annotation suitable for representing written language resources such as corpora. Its goal is to unify a variety of linguistic annotations in one single rich format, using a generic paradigm and without committing to any particular standard annotation set. Instead, it seeks to accommodate any desired system or tagset, and so offer maximum flexibility. This makes FoLiA language independent. Due to its generalized set up, it is easy to extend the FoLiA format to suit your custom needs for linguistic annotation. Using FoLiA for SoNaR offers the flexibility for later projects to enrich SoNaR documents with other kinds of linguistic annotation not provided in the SoNaR project, without needing to switch formats.

FoLiA inherits some properties of the D-Coi format, most notably the structural elements, for divisions, paragraphs, sentence, words and other, are equal or similar. However, because of the introduction of a broader paradigm, FoLiA is not backwards-compatible with D-Coi, i.e. validators for D-Coi will not accept FoLiA XML. For backward compatibility, SoNaR-500 will be delivered in both the FoLiA and D-Coi formats. It is however always easy to convert FoLiA to less verbose formats such as D-Coi.

FoLiA offers an extensive framework for linguistic annotation, all linguistic annotations are implemented as XML elements (as opposed to XML attributes as in D-Coi), and a combination of in-line and stand-off annotation is used to accommodate various linguistic annotation types. The global structure of a FoLiA document is as illustrated below:

```
<FoLiA>
        <metadata src= metadata.cmdi   type=
              cmdi  >
                 <annotations>
                         [declaration of annotation
                            types occurring in the
                            document]
                 </annotations>
        </metadata>
```

```
        <text>
[body of the text to be annotated]
        </text>
</FoLiA>
```

In the metadata block, a reference is made to an external CMDI file which holds the metadata. Furthermore, a mandatory annotations block describes which annotation types are present in the document and declares what tagset is used per annotation type.

An example of the text body in FoLiA XML could look like this, this can be compared with the same excerpt in D-Coi XML in the previous section.

```
<div xml:id="WR-P-E-J-0000000001.div0.1">
   <head xml:id="WR-P-E-J-0000000001.head.1">
      <s xml:id="WR-P-E-J-0000000001.head.1.s.1">
        <w xml:id="WR-P-E-J-0000000001.head.1.s.1.w.1">
          <t>Stemma</t>
          <pos class="N(soort,ev,basis,onz,stan)"/>
          <lemma class="stemma"/>
        </w>
      </s>
    </head>
    <p xml:id="WR-P-E-J-0000000001.p.1">
      <s xml:id="WR-P-E-J-0000000001.p.1.s.1">
        <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.1">
          <t>Stemma</t>
          <pos class="N(eigen,ev,basis,zijd,stan)" />
          <lemma class="Stemma" />
        </w>
        <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.2">
          <t>is</t>
          <pos class="WW(pv,tgw,ev)"/>
          <lemma class="zijn"/>
        </w>
        <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.3">
          <t>een</t>
          <pos class="LID(onbep,stan,agr)"/>
          <lemma class="een"/>
        </w>
        <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.4">
          <t>ander</t>
          <pos class="ADJ(prenom,basis,zonder)"/>
```

```
      <lemma class="ander" />
    </w>
    <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.5">
      <t>woord</t>
      <pos class="N(soort,ev,basis,onz,stan)" />
      <lemma class="woord" />
    </w>
    <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.6">
      <t>voor</t>
      <pos class="VZ(init)" />
      <lemma class="voor" />
    </w>
    <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.7">
      <t>stamboom</t>
      <pos class="N(soort,ev,basis,zijd,stan)" />
      <lemma class="stamboom" />
    </w>
    <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.8">
      <t>.</t>
      <pos class="LET()" />
      <lemma class="." />
    </w>
  </s>
</p>
```

However, the above example is still fairly limited; behind what is shown here is a generalized paradigm applicable to many more types of linguistic annotation. A full overview is beyond the scope of the SoNaR documentation, and for this we refer to the FoLiA documentation itself, to be found at `http://ilk.uvt.nl/folia/`

A RelaxNG validator for FoLiA is also provided there, or via the direct link `https://github.com/proycon/folia/blob/master/schemas/folia.rng`

In SONAR-500 each directory at text category level with FoLiA files also contains an XSLT scheme file (sonar-foliaviewer.xsl). This file allows the user to open a FoLia XML file in a browser with a user-friendly view (without XML-codes).

On the FoLiA website also a number of tools and libraries are provided for working with FoLiA, mostly geared towards developers. There is a Python Library for the reading, creation, search, and manipulation of FoLiA documents from the Python programming language, and there is also a C++ library available (libfolia). Various converters are also provided, such as to

the D-Coi format or to xhtml/css for instant visualisation of a FoLiA document in your browser.

4.2 SoNaR-1 File Formats

The annotations for the SoNaR-1 corpus are currently only 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.

Chapter 5

SoNaR-500 Linguistic Annotations

5.1 Introduction

5.2 Normalization and correction

Where diacritics were missing and the word form without diacritics was 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 Woordenlijst Nederlandse Taal [35] and [36], i.e. the official Word list of the Dutch Language. Also, on the basis of a list of about 16,500 known typos for Dutch most of the texts in the corpus have been screened for these.

Text correction was performed by 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 was 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. The method is described in more detail in [21].

5.3 Language recognition

Where deemed necessary or desirable during processing, the TextCat[1] tool for language recognition was applied. Depending on the source and origin

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

of the texts this was variously done 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 categorization was included in the metadata.

5.4 Corpus annotation

Except for data from the social media (chat, twitter, and SMS), the SoNaR-500 corpus was tokenized by means of ILKTOK, automatically annotated for part of speech and lemmatized by means of FROG [3], while the data was also annotated for named entities by means of Nerd [8].

5.4.1 Part-of-speech tagging and lemmatization

For the tagging and lemmatization of the reference corpus we aimed to yield annotations that were compatible to those in the CGN project. 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. Moreover, some tags that already existed in the original CGN tag set in the D-Coi/SoNaR version cover additional phenomena. The tag set is documented in [33].

Parts of the corpus were tagged at an earlier stage, either in the DCOI project or in the early stages of the SONAR project. For data tagged in the DCOI project the source of error is the human annotator who manually verified the tagging. Next, in the early stages of the SONAR project, when the DCOI tagger-lemmatizer was no longer available, we used its successor Tadpole. Later on, Tadpole was replaced by FROG[2]. Therefore, towards the end of the project we used FROG to tag the bulk of the corpus.

[2]FROG is available under GPL (online demo: `http://ilk.uvt.nl/cgntagger/`, software: `http://ilk.uvt.nl/frog/`

In order to overcome the imperfections and inconsistencies of the part-of-speech tagging performed over the course of the years the SoNaR project ran, we have in a recent curation phase redone the part-of-speech tagging and lemmatization with a single, consistent recent version of FROG. The legacy annotations are still available for backwards compatibility reasons, but we advise all users to use only the annotations labeled: "frog-mbpos-1.0" for the POS-tagging, "frog-mblem-1.1" for the lemmatization and "frog-mbma-1.0" for the morphological analyses.

If you have a version of SoNaR in which the FoLiA xml files do not have the following lines in the metadata header, you have an older, uncurated version of the corpus and we strongly advise you to apply to the Ducth HLT agency, TST-Centrale, for the curated release.

Metadata header lines:

```
<pos-annotation annotator="frog-mbpos-1.0"
    annotatortype="auto" datetime="
2013-02-14T22:24:20" set="http://ilk.uvt.nl/folia/sets/
    frog-mbpos-cgn" />
<lemma-annotation annotator="frog" annotatortype=
    "auto" set="hdl:1839/00-S
CHM-0000-0000-000E-3" />
<lemma-annotation annotator="frog-mblem-1.1"
    annotatortype="auto" datetime
="2013-02-14T22:24:20" set="http://ilk.uvt.nl/folia/
    sets/frog-mblem-nl" />
<morphological-annotation annotator="frog-mbma
    -1.0" annotatortype="auto" datetime="
    2013-02-14T22:24:20" set="http://ilk.uvt.nl/
    folia/sets/frog-mbma-nl" />
```

5.4.2 Named entity annotation

For the annotation of named entities in SoNaR, a new annotation scheme was developed for Dutch. The intuition and motivation for this scheme is discussed in [8]. In brief, the annotation scheme allows for the annotation of six main named entity types, subtype annotation, and the annotation of metonymic usage. The six main named entity types are the following:

1. Persons (PER)
2. Organizations (ORG)

3. Locations (LOC)

4. Products (PRO)

5. Events (EVE)

6. Miscellaneous named entities (MISC)

The Named Entity Recognition for Dutch (NERD) classifier is available in the NERD directory (cf. Section 3.2.4).

It does named entity recognition of the 6 main types, and returns IOB-formatted output.

The classifier uses CRF++ for classification, and the following 22 features:

- basic features: token, part-of-speech-tag (CGN tagset)
- morphological features (binary): firstCap, allCaps, internalCaps, allLowercase, containsDigit, containsDigitAndAlpha, onlyDigits, isPunctuation, containsPunctuation, isHyphenated
- regular expressions (binary): does the token match regular expression describing initials or URLs?
- word length in characters
- prefix and suffix information: character n-grams of length 3 and 4 at the beginning and end of the token
- function word (binary): is the token present in a predefined list of function words?
- first token (binary): is this the first token of a sentence?
- word shape (discrete): to which form of a predefined set of word forms does the token belong?

The classifier was trained on the manual annotations of the entire SONAR-1 corpus.

5.4.3 Morphological analysis

During the final curation phase of SoNaR-500 we decided to add an extra layer of linguistic annotation, i.e. morphological analysis. This was performed by MBMA which is a part of Frog.

Chapter 6

Metadata

6.1 Introduction

Metadata provided with the corpus relates to the text itself (e.g. Text type, Collection name, Place and date of publication). Also included is information pertaining to the type of IPR license that applies. IPR has been arranged for (nearly) all data. How this was achieved is described in [7].

IPR License Codes for texts acquired in the Netherlands or Belgium typically have the following format:

SoNaR.< 1-7> . < NL,VL> -< A,B,C> .< nr>

- Texts acquired within the pilot project D-Coi may have D-Coi/SoNaR as their first element.
- The numerical codes 1-7 detailing the types of IPR agreements that were used are described in Table 6.1.
- The second numerical code, preceded by a hyphen, details whether or not restrictions on use of the text for commercial purposes apply to the particular text. Whether or not the text has IPR restrictions is encoded by:
  - -0 : no restriction on use for commercial purposes applies to the text
  - -1 : restrictions on use for commercial purposes apply to the text
- NL, VL indicate if the text was acquired in The Netherlands (NL) or Flanders (VL).

  - A refers to acquisition done by Twente University
  - B refers to acquisition done by Tilburg Unversity
  - C refers to acquisition done by Radboud University Nijmegen

- nr is the number assigned to the license in the respective category.

Not all elements are necessarily contained in every license code.

Every IPR agreement has been assigned a code, which encodes a lot of valuable metadata about the agreement/donation. Thus each snippet of text can be traced back to its origin.

For example: the license agreement code SoNaR.2-1.NL-B.00015 indicates that this is an agreement reached within SoNaR, that a standard agreement for publishers (2) has restrictions ('-1': No agreement concerning commercial use, i.e. texts were donated for research purposes only)

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

All metadata is contained in so called CMDI files, which we describe in the next section. The files in D-Coi+ format do not contain any metadata. CMDI metadata files are currently only available for SoNaR-500, not for SoNaR-1.

6.2 Metadata format: CMDI

For registering the metadata in SoNaR, we chose to adopt the CMDI format. CMDI is dynamical and it fits well in the CLARIN infrastructure. The CMDI structure enables the provider of the corpus to add or remove metadata categories in the future. For this purpose, a profile named SoNaRcorpus was composed for the SoNaR corpus. In this profile existing ISOcat elements and adaptations of existing CMDI components were combined with newly created components and elements.

A profile contains several components. Components can contain components and/or elements. A component or an element can occur a minimum and a maximum number of times, depending on the limitations as defined in the profile. As for the SoNaR corpus profile, one element (Annotation-Type-SoNaR) may contain one of seven different annotation types, used in SoNaR:

- token

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 6.1: Types of IPR-agreements used in SoNaR

- lemma [3]
- POS [3]
- named entity [8]
- coreference relations [6] and [11] and [12]
- semantic roles [32] and [16]
- spatiotemporal relations [25] and [26]

The maximum number of times this field can be added, is defined as unbounded.

In fact in the current version of SoNaR we have two distinct lemma and POS annotation layers. The first was the original lemmatisation (identified by reference 'hdl:1839/00-SCHM-0000-0000-000E-3') and POS-tagging (identified by reference 'hdl:1839/00-SCHM-0000-0000-000B-9') added throughout the years the corpus was built. In so far as the tool used underwent a gradual transformation as it was being further developed and improved from Tadpole to its current incarnation Frog these original layers are of unequal quaility and consistency. To remedy this, after the project ended and both internal and external evaluations had brought to light certain shortcomings, a new and uniform lemmatization and POS-tagging layer was added to the whole of the corpus. The original layers ensure compatibility with the results of the Lassy and DutchSemCor projects. The new layers are recommended to corpus users who have no need for this backward compatibility, in the latter curation phase SoNaR underwent several million known errors were remedied. These newer annotations are identified by the references:

- POS: pos-annotation annotator="frog-mbpos-1.0"
- Lemmatization: lemma-annotation annotator="frog-mblem-1.1"
- Morphological analysis: morphological-annotation annotator="frog-mbma-1.0"

In the following Section we give the complete list of components available in the SoNaRcorpus CMDI profile. The structure of this overview is as follows:

<ComponentName>
    [Description of component]

6.2.1 Complete list of components available in the SoNaR corpus CMDI profile

```
<Text>
<ProjectName>
     [Majority of the texts result from the SoNaR
        project]

<CollectionName>
     [Name of a collection the text belongs to]
<CollectionCode>
     [Administrative code for collection of text]
<CollectionDescription>
     [Description of the collection text belongs to]

<TextTitle>
     [Title of the text]
<TextSubTitle>
     [Subtitle of the text]
<TextIntro>
     [Short introduction or abstract pertaining to
        the text]
<TextDescription>
     [Description of the text]
<TextType></TextType>
     [Category of the text]
<TextClass>
     [General domain information (if at all
        available)]
<TextKeyword>
     [Keywords pertaining to the text]
 <TotalSize>
 <Number>
     [Number of units]
 <SizeUnit>
     [Definition of unit, i.e. tokens, bytes etc.]

 <Language>
    <LanguageName>
    <ISO639>
```

```
   <iso-639-3-code>
    [Language of text according to ISO 639-3]

<License>
   <LicenseCode>
    [Administrative code, ascribed to license.]
   <LicenseType>
    [Type of IPR license, cf. Table~\ref{IPR}.]
   <LicenseDetails>
    [Details or exceptions on above named license.]
   <LicenseDate>
    [License conclusion s date.]

<Source>
  <SourceName>
    [Name of the source of text, e.g. name of
       publisher s house that donated a text. For
        new media it can contain the source chatbox
        of chats or the donation channel for SMS.]
 <Continent>
    [Continent of origin of the text]
 <Country>
    [Country of origin of the text]
 <OutputTextcat>
    [Output of language classification software
       TextCat for the text]
 <SourceLanguageIdentifications>
    [Language of source of text]
 <OriginalFile>*
   <OriginalFileName>
    [Name of original file]
   <OriginalFileDate>
    [Date of original file]
   <OriginalFileAcquisitionDate>
    [Date of acquisition of original file.]
   <TotalSize>
      <Number>
    [Number of units]
      <SizeUnit>
    [Definition of unit, i.e. tokens, bytes etc. of
       original file.]
```

```
<Publication>
   <Published>
   [Indicates if a text was published or not]
   <Publisher>
   [Name of publishing house]
   <PublicationName>
   [Name of publication]
   <PublicationPlace>
   [Place of publication]
   <PublicationDate>
   [Date of publication]
   <PublicationTime>
   [Time of publication]
   <PublicationVolume>
           [Volume of edition in which text was
              published]
   <PublicationNumber>
   [Number of issue]
   <PublicationIssue>
   [Issue of publication]
   <PublicationSection>
   [Section of edition/issue that contained the
      text]
   <PublicationSubSection>
   [Subsection of edition/issue that contained the
       text.]
   <PublicationPage>
   [Page number on which text was published]
   <PublicationId>
   [Code ascribed for publication, e.g. at
      publisher s house]
   <PublicationLanguage>
   [Language of text, as named in original file.]
    <PublicationScope>
   [Scope of the publication]
    <PublicationGenre>
   [Type of movie or TV serial (typical for movie
      subtitles)]

 <BroadcastPublication>
```

```
   <BroadcastId>
   <BroadcastDate>
   <BroadcastTotalDuration>
   <BroadcastAudioDuration>
   <BroadcastAirtime>
   <BroadcastPresenter>
   <BroadcastModificationDate>
   <BroadcastModifier>

 <Author>
    <Name>**
  [Name of author]
    <Pseudonym>
  [Pseudonym of author. For new media this field
     is used for 'nickname' in for example
     chatbox]
   <Sex>
  [Gender of author]
   <Age>
  [Age of author]
   <MotherTongue>
  [Mother tongue of author]
   <ResidencePlace>
  [Place of residence of author]
   <Country>
  [Country of residence of author]
   <Town>
  [Town of residence of author]

<Translation>
   <Translated>
  [Whether a text is translated or not]
   <TranslatorName>
  [Name of translator]
   <OriginalLanguage>
   [Original language before translation]

 <Annotationtypes-SoNaR>
    <AnnotationType-SoNaR></AnnotationType-SoNaR>
  [Levels of annotations:
                - token
```

- lemma
- POS
- morphological analysis
- named entity

(SoNaR-1 only:
- coreference relations
- semantic roles
- spatiotemporal relations)]

Elements with an asterix (*) are not provided to the users of the corpus. Elements with double asterix (**) are not provided to the users of the corpus in the case of chats, tweets and SMS.

All components are incorporated in a component called Text and this component is included in the ISOcat profile SoNaRcorpus.

```
<Language ComponentId="clarin.eu:cr1:c_1271859438111">
          <LanguageName>Dutch</LanguageName>
          <ISO639 ComponentId="clarin.
              eu:cr1:c_1271859438110">
            <iso-639-3-code>nld</iso-639-3-code>
          </ISO639>
        </Language>
```

If known, the country of origin is also stated in the field:

```
<Country>NL/B</Country>
```

Below an example is given of a metadata scheme according to this CMDI profile.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <CMD xmlns="http://www.clarin.eu/cmd/" xmlns:xsi="
    http://www.w3.org/2001/XMLSchema-instance"
    CMDVersion="1.1" xsi:schemaLocation="http://www.
    clarin.eu/cmd/ http://catalog.clarin.eu/ds/
    ComponentRegistry/rest/registry/profiles/clarin.
    eu:cr1:p_1328259700943/xsd">
<Header />
- <Resources>
<ResourceProxyList />
<JournalFileProxyList />
<ResourceRelationList />
</Resources>
```

```
- <Components>
- <SoNaRcorpus>
- <Text ComponentId="clarin.eu:cr1:c_1328259700942">
<ProjectName>SoNaR</ProjectName>
<CollectionName>Books</CollectionName>
<CollectionCode>Book published by Uitgeverij Maarten
    Muntinga</CollectionCode>
<CollectionDescription />
<TextTitle>Stille wateren</TextTitle>
<TextSubTitle />
<TextIntro />
<TextDescription>Information collected from the book's
    title pages: Stille wateren | Barbara Nadel | Stille
    wateren | Een inspecteurs Ikmen & Suleyman-
    detective | Sirene | Oorspronkelijke titel Deep
    waters | Oorspronkelijk verschenen bij Headline Book
    Publishing, London |  2002 Barbara Nadel |  
    2004 Nederlandse vertaling Uitgeverij Sirene,
    Amsterdam | Vertaald door Noor Koch | Omslagontwerp
    Studio Eric Wondergem BNO | Foto voorzijde omslag
    Corbis | Foto achterzijde Paul's Studio | Zetwerk
    Stand By, Nieuwegein | Druk Bercker, Kevelaer |
    Uitgave in Sirene juni 2004 | Alle rechten
    voorbehouden | Uitgeverij Sirene is een onderdeel
    van Uitgeverij Maarten Muntinga bv | www.sirene.nl |
     isbn 90 5831 309 3 | nur 331 | ||| Online book
    description (in Dutch): Het inspecteursduo Cetin
    Ikmen en Mehmet Suleyman moet in Istanbul de moord
    op Rifat Berisha oplossen. Zijn familie is niet
    geneigd te helpen bij het onderzoek, omdat de
    familie met de daders in "bloedwraak" zijn. |
    Source: URL: http://www.inejacet.nl/
    GerdBoerenrecensies/BarbaraNadelrecensies.htm</
    TextDescription>
<TextType>WR-P-P-B_books</TextType>
<TextClass />
<TextKeywords ComponentId="clarin.
    eu:cr1:c_1328259700939" />
- <TotalSize ComponentId="clarin.eu:cr1:c_1271859438114
    ">
<Number>137670</Number>
```

<SizeUnit>word tokens</SizeUnit>
</TotalSize>
- <Language ComponentId="clarin.eu:cr1:c_1271859438111">
<LanguageName>Dutch</LanguageName>
+ <ISO639 ComponentId="clarin.eu:cr1:c_1271859438110">
<iso-639-3-code>nld</iso-639-3-code>
</ISO639>
</Language>
- <License ComponentId="clarin.eu:cr1:c_1328259700915">
<LicenseCode>SoNaR.2.NL-B.012</LicenseCode>
<LicenseType />
<LicenseDetails>Uitgeverij Maarten Muntinga, http://www.sirene.nl</LicenseDetails>
<LicenseDate />
</License>
- <Source ComponentId="clarin.eu:cr1:c_1328259700940">
<SourceName />
<Continent>Europe</Continent>
<Country>NL</Country>
<SourceLanguageIdentification />
- <Publication ComponentId="clarin.eu:cr1:c_1328259700917">
<Published />
<Publisher>Uitgeverij Maarten Muntinga URL: http://www.sirene.nl/</Publisher>
<PublicationName />
<PublicationPlace />
<PublicationDate>2004-06</PublicationDate>
<PublicationTime />
<PublicationVolume />
<PublicationNumber />
<PublicationIssue />
<PublicationSection />
<PublicationSubSection />
<PublicationPage />
<PublicationId />
<PublicationLanguage />
<PublicationScope />
<PublicationGenre />
</Publication>

```
- <BroadcastPublication ComponentId="clarin.
    eu:cr1:c_1328259700914">
<BroadcastId />
<BroadcastDate />
<BroadcastTotalDuration />
<BroadcastAudioDuration />
<BroadcastAirtime />
<BroadcastPresenter />
<BroadcastModificationDate />
<BroadcastModifier />
</BroadcastPublication>
</Source>
- <Author ComponentId="clarin.eu:cr1:c_1328259700913">
<Name>Nadel, Barbara</Name>
<Pseudonym />
<Sex />
<Age />
<MotherTongue />
- <ResidencePlace ComponentId="clarin.
    eu:cr1:c_1324638957682">
<Country />
<Town />
</ResidencePlace>
</Author>
- <Translation ComponentId="clarin.
    eu:cr1:c_1328259700921">
<Translated>Y</Translated>
<TranslatorName>Noor Koch</TranslatorName>
<OriginalLanguage />
</Translation>
- <Annotationtypes-SoNaR ComponentId="clarin.
    eu:cr1:c_1328259700912">
<AnnotationType-SoNaR>Token</AnnotationType-SoNaR>
<AnnotationType-SoNaR>Lemma</AnnotationType-SoNaR>
<AnnotationType-SoNaR>POS</AnnotationType-SoNaR>
<AnnotationType-SoNaR>NER</AnnotationType-SoNaR>
</Annotationtypes-SoNaR>
</Text>
</SoNaRcorpus>
</Components>
</CMD>
```

6.2.2 Domain information

For at least some large subcollections of particular text types we have domain information. This information is not necessarily only available in the most obvious metadata field, but the field 'TextClass' is the most likely place to find this information. Domain information per text nevertheless is not available for large parts of SoNaR-500 as it was not even available in the original text collections provided to us by our donators.

In the project DutchSemCor the full SoNaR-500 corpus has undergone automatic domain labeling per text. This further enhanced version of the corpus may also become available through the HLT Agency TST-Centrale.

Chapter 7

SoNaR Frequency lists

SONAR-500 comes with various frequency lists. There are word frequency lists, lemma frequency lists, and lemma+POS frequency lists. These are contained in simple tab-separated values or TSV files. These files are provided for each text type and aggregated over the full corpus.

The frequency files contain four TAB delimited columns:

- word OR lemma OR Lemma+Pos
- absolute frequency
- accumulated frequency
- accumulated frequency in rounded percentages

In the case of the frequency files aggregated over all the distinct text types, i.e. over the whole SoNaR-500 corpus, the hapaxes have been omitted due to computer memory restrictions. The hapaxes are those items that have frequency 1 in the distinct frequency lists per text type.

Chapter 8

SoNaR-500: Contents

8.1 Older media

We refer the interested user to the SQL-database or to the CMDI metadatafiles for descriptions of the actual contents of the older media sections of SoNaR-500.

We have subtitles for 1275 films, offering over 9.5MW from OpenSubtitles. These were obtained from the Opus Corpus (URL: `http://opus.lingfil.uu.se/`), who originally obtained these from `http://www.opensubtitles.org/`. We gratefully acknowledge this contribution.

8.2 New media

Texts from the social media have a somewhat special status in the corpus. The texts and metadata were collected from a variety of social media, viz. chat, twitter, SMS. Special issues for this material are:

- the source and collection method of the data;
- the reliability of author information;
- anonymization of texts;
- the occurrence of time stamped events

For these reasons the various texts from the social media included in SONAR-500 are described in more detail below.

Due to the irregularities in spelling typically encountered in these texts, automatic POS tagging, lemmatization and labeling of Named Entities was not at first performed in SoNaR. At the time of refrogging the corpus this

was nevertheless performed, but no guarantees are given nor claims made as to the quality of these automatic annotations. Note that for these media only a single, new, layer of lemmatization and pos-tagging annotations are available. Named entity labeling is still not available.

8.2.1 Chat

By chats we mean real time typed conversations over a computer network between two or more people.

With chats three possible events are possible:

1. a user enters a message (concluded by pressing the send button or typing the enter key)

2. a user joins or leaves a chat room (collection of users who can see each other's messages)

3. a user changes his/her nick name (users often use an invented name, other than their own)

In general, each event is represented with a date and time stamp, the nick name of the user and in case of a message, the content of the message.

The chats were collected from different sources. For some sources, date and time stamps are not available or imprecise as well as information about joining/leaving the chat room.

Messages have been tokenized by means of UCTO . Counts of the number of tokens, as presented in Table 8.1, are based upon the tokenization.

There is a clear distinction in the chat data from Dutch users and from Flemish users. The data collected in the Netherlands come from four different sources, described below. Of all these chats, the users gave individually permission for the data to be republished (though not always with SoNaR in mind). Of most users metadata (age, sex and residence) are known. Although a nick name in chat seldom reflects the users real identity, all nick names have been anonymized, both in the field that indicates the sender of the message, as in the messages themselves. No further anonymization has been done of e.g. other names, addresses, telephone numbers. In an internal study carried out to investigate the possibilities of (automatic) anonymization, this seemed not feasible.

Table 8.1 shows the number of word tokens for the different gender and age categories for all chat data of Dutch users. [1]

[1]For the Flemish (BEL) data no metadata are available. In case no (reliable) information was available for Sex and/or Age the data have been listed as Unknown (Unkn.)

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	91,776	12.44	127,404	17.27	–	–	219,180	29.72
21-40	69,333	9.40	178,739	24.24	–	–	248,072	33.64
41-60	138,690	18.80	6,308	0.86	–	–	144,998	19.66
61-99	–	–	–	–	–	–	–	–
Unkn.	39,763	5.39	55,718	7.55	29,789	4.04	125,270	16.99
total	339,562	46.04	368,169	49.92	29,789	4.04	737,520	100.00

Table 8.1: Composition of the Dutch (NLD) chat data. The number of words and percentage per age/sex category of all Dutch chats

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	–	–	–	–	–	–	–	–
21-40	34,412	9.73	178,739	50.56	–	–	213,151	60.29
41-60	134,378	38.01	5,932	1.68	–	–	140,310	39.69
61-99	–	–	–	–	–	–	–	–
Unkn.	–	–	–	–	80	0.02	-80	0.02
total	168,790	47.74	184,671	52.23	80	0.02	353,541	100.00

Table 8.2: Composition of the Dutch #lands chat data. The number of words and percentage per age/sex category of all #lands chats

What follows is a more detailed description of the different sources of the chat data.

#lands Staff from the linguistics department at the Radboud University Nijmegen used a chatbox that was set up for SoNaR, after the coffee break in the morning. Data was collected between December 8, 2010 and February 17, 2012. During this period, a reminder e-mail was sent to the participants on each workday. After entering the chatbox, a statement was shown, to make clear that the data in this chatbox would be made available for the SoNaR corpus. The chatbox was used by 30 participants. From all users gender, age and country and town of residence were registered. The size of this subcorpus is 353,541 word tokens. Its composition is given in Table 8.2.

What follows is a more detailed description of the different sources of the chat data.

Unknown (Unkn.) data result from a chatter who disguised his/her identity by changing nicknames several times.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	30,415	36.29	25,227	30.10	–	–	55,642	66.39
21-40	–	–	–	–	–	–	–	–
41-60	–	–	376	0.45	–	–	–	–
61-99	–	–	–	–	–	–	–	–
Unkn.	712	0.85	534	0.64	26.542	31.67	376	0.45
total	31,127	37.14	26,137	31.19	26,542	31.67	83,806	100.00

Table 8.3: Composition of the Dutch #ChatIG chat data. The number of words and percentage per age/sex category of all #ChatIG chats. (For the 2nd year no metadata are available.)

#chatig The chatIG corpus is a corpus of chat language of Dutch teenagers. It was collected by Wilbert Spooren between 2004 and 2006, i.e. prior to the SoNaR project. Different classes from secondary schools in Amsterdam came to the VU to chat via the chat room tool in Blackboard. These chat conversations were regulated and topics were provided. The groups consisted of boys only or girls only, while some groups were mixed. After the sessions the pupils filled in a survey, in which they provided demographical information and information about their use of social media. Because of the demographics it was possible to set up a scheme which matched students who were intimate with each other and students who are not. (Students were asked to fill in the names of approximately five other students from their class whom they spent the most time with.)

The parents of the pupils gave permission for the chats to be published.

For the data of 2004-2005 a database with metadata was created. It contains information about the participating pupils, their classes, the chat sessions and the subjects of the chat sessions. Of this information, only age, gender and residence have been included in the SoNaR metadata. For the years 2005-2006 and 2007 no metadata are available.

The size of this subcorpus is 83,806 word tokens. Its composition is given in Table 8.3.

For the sessions from 2004-2005; 2005-2006 a detailed description can be found in Charldorp (2005).

The task used in the 2007 sessions is described in Spooren (2009).

#bonhoeffer The Bonhoeffer subcorpus was collected at a secondary school in Enschede, The Netherlands, in April 2010. In cooperation with the teach-

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	8.780	31.43	16,181	57.92	–	–	24,961	89.35
21-40	–	–	–	–	–	–	–	–
41-60	–	–	–	–	–	–	–	–
61-99	–	–	–	–	–	–	–	–
Unkn.	–	–	–	–	2,975	10.65	–	–
total	8.780	31.43	16,181	57.92	2,975	10.65	27,936	100.00

Table 8.4: Composition of the Dutch #bonhoeffer chat data. The number of words and percentage per age/sex category of all #bonhoeffer chats.

ers, a class of students was divided in groups of four students, who had chat conversations in a chatbox, set up for SoNaR. In each group, there were two chat sessions with two persons participating and one chat session with four students. Each session lasted 10 minutes and topics for the chats were provided, although students were allowed to chose their own topic as well.

The parents of the students gave permission for the chats to be published.

The size of this subcorpus is 27,936 word tokens. Its composition is given in Table 8.4.

Unknown data result from a couple of chatters who disguised their identities by changing nicknames several times.

More information about the Bonhoeffer chat corpus can be found in Uittenhout (2010).

#msn A total of 1,056 chat sessions (64,116 lines) of different kinds, including for example MSN and organized chat sessions were collected by Remy van Rijswijk and Sonja van der Hoek in the framework of the NEWSPEAK project . The collection of the data took place between October 2009 and April 2010.

Recruitment was organized through a chain letter sent to friends and family. All participants signed a form in which they gave permission to use the data for scientific research. The same form was used to add metadata to the corpus, like age, gender and region of birth. 452 sessions were included in the SoNaR corpus.

The size of this subcorpus is 272,237 word tokens. Its composition is given in Table 8.5.

Unknown data result from a chatter who disguised his/her identity by changing nicknames several times.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	52.581	19.31	85,996	31.59	–	–	138,577	50.90
21-40	34,921	12.83	–	–	–	–	34,921	12.83
41-60	4,312	1.58	–	–	–	–	4,312	1.58
61-99	–	–	–	–	–	–	–	–
Unkn.	39.051	14.34	55,184	20.27	192	0.07	94,427	34.69
total	130,865	48.07	141180	51.86	192	0.07	272,237	100.00

Table 8.5: Composition of the Dutch #msn chat data. The number of words and percentage per age/sex category of all #msn chats.

#chat.be The Flemish website www.chat.be gave permission to use chats from its website. Chats were (not continuously) collected between March 4, 2011 and February 11, 2012. The chats are from the main chat channel of the site (named chat.be). Participants did not give permission individually and no metadata of the participants is available. No anonymization of the data has been carried out.

The size of this subcorpus is 11,135,664 word tokens.

8.2.2 Twitter

Tweets are messages published via twitter.com. Only tweets that were publicly available are collected. The Guidelines for Use of Tweets in Broadcast or Other Offline Media [2] state that it is allowed to republish tweets, but only unchanged. Therefore the tweets have not been anonymized or altered in any other way.

Some twitterers publish tweets both in Dutch and in another language, primarily English. Twitterers who only publish in non-Dutch were removed from the corpus, but no language detection was done to remove single tweets that are non-Dutch. As a consequence quite a few English tweets ended up in the SoNaR corpus.

All tweets have been tokenized with UCTO. Counts of the number of tokens, as presented in Table 8.6, are based upon the tokenization.

Table 8.6 shows the number of word tokens for the different gender and age categories for all twitter data.

The tweets in the SoNaR corpus can be divided in two subcorpora: Submitted and Found. A description of each of these can be found below.

[2] `https://support.twitter.com/entries/114233`

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	1,790,502	7.72	664,354	2.86	–	–	2,454,856	10.58
21-40	8,011,061	34.53	4,698,220	20.25	–	–	1,270,928	54.79
41-60	4,031,100	17.38	2,295,901	9.90	–	–	6,327,001	27.27
61-99	188,069	0.81	202,856	0.87	–	–	390,925	1.69
Unkn.	887,470	3.83	427,678	1.84	–	–	1,315,148	5.67
total	14,908,202	64.27	8,289,009	35.73	–	–	23,197,211	100.00

Table 8.6: Composition of the twitter subcorpus. The number of words and percentage per age/sex category of all tweets.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	1,729,492	10.35	663,471	3.97	–	–	2,392,963	14.32
21-40	5,837,721	34.94	3,500,732	20.96	–	–	9,338,453	55.90
41-60	2,682,420	16.06	1,906,634	11.41	–	–	4,589,054	27.47
61-99	182,392	1.09	202,856	1.21	–	–	385,248	2.31
Unkn.	–	–	–	–	–	–	..	–
total	10,432,025	62.45	6,273,693	37.55	–	–	16,705,718	100.00

Table 8.7: Composition of the submitted tweets subcorpus. The number of words and percentage per age/sex category of all submitted tweets.

Submitted The subcorpus of submitted tweets contains (mainly) Dutch tweets. A tweet about the tweet collection in SoNaR with a request for metadata of twitterers caused a snowball effect. Attention for the tweet collection was spread over twitter, and also a national news website, and the Radboud Universitys homepage reported on the data collection. Twitterers were asked to participate and to submit the name of their Twitter account and metadata to the SoNaR team, either by email of via a webform.

The Twitter API [3] was used to collect the tweets for the corpus. Retweets were not collected. As for the metadata: gender, age and town of residence of the twitterer are available in the corpus.

The size of this subcorpus is 16,705,718 word tokens. Its composition is given in Table 8.7.

[3] http://api.twitter.com

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	61,010	0.94	883	0.01	–	–	61,893	0.95
21-40	2,173,340	33.48	1,197,488	18.45	–	–	3,370,828	51.93
41-60	1,348,680	20.78	389,267	6.00	–	–	1,737,947	26.77
61-99	5,677	0.09	–	–	–	–	5,677	0.09
Unkn.	887,470	13.67	427,678	6.59	–	–	1,315,148	20.26
total	4,476,177	68.95	2,015,316	31.05	–	–	6,491,493	100.00

Table 8.8: Composition of the found tweets subcorpus. The number of words and percentage per age/sex category of all found tweets.

Found The subcorpus of found [4] tweets contains tweets from Dutch and Flemish (semi-)celebrities, such as politicians, actors and sports people. From public websites such as Wikipedia and home- or fan pages the corresponding metadata (gender, age or birth date and either birth town or town of residence) were collected.

The size of this subcorpus is 6,491,493 word tokens. Its composition is given in Table 8.8.

8.2.3 SMS

The SMS corpus in SoNaR is a collection of Short Message Service messages, collected in Flanders and The Netherlands between September and December, 2011. Only messages that were sent by the contributor are included in the corpus.

One iPad per country was put up for raffle among all contributors of SMS texts to the corpus.

Text messages are tokenized with UCTO. Counts of the number of tokens, as presented in Table 8.9, are based upon the tokenization.

Table 8.9 shows the number of word tokens for the different gender and age categories for all SMS data.

Technically, contributors delivered SMS messages to the corpus through three different channels. The delivery channel results in slightly different properties of the data in terms of transcription and anonymization. Relevant remarks will be explained below.

[4]Found: these tweets were not submitted by the creators but collected by the SoNaR team by crawling the twitter pages.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	33,000	4.56	11,036	1.52	–	–	44,036	6.08
21-40	271,819	37.55	196,867	27.20	–	–	468,686	64.75
41-60	10,226	1.41	1,867	0.26	–	–	12,093	1.67
61-99	–	–	63	0.01	–	–	63	0.01
Unkn.	169,737	23.45	24,803	3.43	4,458	0.62	198,998	27.49
total	484,782	66.97	234,636	32.41	4,458	0.62	723,876	100.00

Table 8.9: Composition of the SMS subcorpus. The number of words and percentage per age/sex category of all SMS texts.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	1,223	6.12	6,856	34.33	–	–	8,079	40.46
21-40	2,028	10.16	7,337	36.74	–	–	9,365	46.90
41-60	337	1.69	1,820	9.11	–	–	2,157	10.80
61-99	–	–	63	0.32	–	–	63	0.32
Unkn.	–	–	305	1.53	–	–	305	1.53
total	3,588	17.97	16,381	82.03	–	–	19,969	100.00

Table 8.10: Composition of the SMS subcorpus of text messages submitted via the online form. The number of words and percentage per age/sex category of all such SMS texts.

Online submission form On the SoNaR website, an online submission form was available in which text messages could by copied manually. Contributors were asked to transcribe (a maximum of) six SMS messages from their SMS outbox (containing only SMS sent by the owner of that device) and to enter their gender, age, country and town of residence. After submitting the form, a box was showed saying click here if you want to add more messages. By clicking the box, the contributor was sent back to the submission form again. No automatic anonymization was done, but selecting appropriate messages was left to the contributor. No time and date stamp are available for the original text messages and neither is information available about the original recipients identity.

The size of this subcorpus is 19,969 word tokens . Its composition is given in Table 8.10.

Android application An application could be installed on mobile devices running on Android, that uploaded the messages from the SMS outbox to a

Original content	Example	Replacement code
Email address	name@gmail.com	(EMAIL)
URL	http://www.google.com	(URL)
IP Address	127.0.0.1	(IP)
Time	12:30	(TIME)
Date	19/01/2011	(DATE)
Decimal	21.3	(DECIMAL)
Integer over 7 digits long	40000000	(#)
Hyphen-Delimited number	12-4234-212	(#)
Alphanumeric Number	U2003322X	U(#)X

Table 8.11: Replacement codes used in anonymizing SMS texts

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	31,777	8.55	4,180	1.12	–	–	35,957	9.67
21-40	147,904	39.78	124,518	33.49	–	–	272,422	73.27
41-60	9,889	2.66	47	0.01	–	–	9,936	2.67
61-99	–	–	–		–	–	–	
Unkn.	53,477	14.38	–		–	–	53,477	14.38
total	243,047	65.37	128,745	34.63	–	–	371,792	100.00

Table 8.12: Composition of the SMS subcorpus of text messages submitted via android app. The number of words and percentage per age/sex category of all such SMS texts.

draft mail in contributors' Gmail account. In this way contributors were given the opportunity to modify or remove text messages before sending out the e-mail. From there the text messages were sent to a dedicated SoNaR mailbox. A time and date stamp is available for each original sent message, as well as a unique identifier, replacing the original recipients phone number. Automatic anonymization was done by the app to replace sensitive data including dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, street numbers, etc.), email addresses, URLs, and IP addresses. Such information was captured using regular expressions and replaced by the corresponding semantic placeholders, as shown in Table 8.11. For example, any detected email address was replaced by the code (EMAIL).

The size of this subcorpus is 371,792 word tokens. Its composition is given in Table 8.12.

Sex	Male		Female		Unknown		Total	
Age	# wrds	%	# wrds	%	# wrds	%	# wrds	%
0-20	–	–	–	–	–	–	–	–
21-40	121,887	36.70	65,012	19.58	–	–	186,899	56.28
41-60	–	–	–	–	–	–	–	–
61-99	–	–	–		–	–	–	
Unkn.	116,260	35.01	24,498	7.38				
total	238,147	71.71	89,510	26.95	4,458	1.34	332,115	100.00

Table 8.13: Composition of the SMS subcorpus of text messages submitted via export files. The number of words and percentage per age/sex category of all such SMS texts.

Export file Different files were generated by dedicated software belonging to the mobile device and sent to the SoNaR mailbox by the contributor. Contributors were able to remove SMS messages before sending the file to SoNaR. Contributors were asked not to modify the messages, but it cannot be guaranteed that contributors did not change the content of the messages. Contributors were told they could replace their own personal first name by a common first name, as a way of anonymizing the data. No (automatic) anonymization was done for the export files. The text files with text messages came in different formats. In some files, time and date info was available, in others it was not.

The size of this subcorpus is 332,115 word tokens. Its composition is given in Table 8.13.

Chapter 9

Beyond SoNaR

The above manual describes the SoNaR corpora SoNaR-500 and SoNaR-1 in detail.

Both today come as large collections of separate files, in separate archives according to the various text categories incorporated in them.

Searching for desirable information, extracting the relevant files and getting the desired information out today requires hard- and software not really available to even most researchers in the Humanities.

To alleviate this problem, CLARIN-NL has recently approved a new project called OpenSoNaR.

The OpenSoNaR project will provide end users with the online means for extracting useful information from the SoNaR-500 reference corpus of contemporary written Dutch. This includes exploring the texts and navigating through the SoNaR-500 corpus by way of the metadata. The project makes the contents of the new SoNaR-500 reference corpus available to layman and specialist researcher alike. Based on the desiderata of four distinct CLARIN-NL priority groups of humanist researchers, access to the corpus for navigation, exploration and exploitation in an online environment will be through a front-end, to be called WhiteLab, providing a range of interfaces that provide user-driven functionality. The backend is INLs new retrieval engine BlackLab, designed to provide access to corpora for linguistic and lexicographical use in the CLARIN infrastructure.

Work in the OpenSoNaR project will be undertaken in close cooperation with the Dutch NWO 'Groot' project Nederlab which aims at eventually making available online all extant Dutch text collections, also providing researchers all necessary means and tools for conducting valuable and exciting new research based on these textual riches, this immense wealth of words.

Bibliography

[1] Beinema P.: A Reference Corpus of Written Dutch. Technical specification of file formats and validation tools used. [1] Technical report TR C-COI-06-02 (2006)

[2] Aston, G., Burnard, L.: The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh (1998)

[3] Van den Bosch, A., Busser, B., Canisius, S., Daelemans, W.: An efficient memory-based morphosyntactic tagger and parser for Dutch. In: P. Dirix, I. Schuurman, V. Vandeghinste, F. Van Eynde (eds.) Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting, pp. 99–114. Leuven, Belgium (2007)

[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 (2002). Nederlandse Taalunie, The Hague, The Netherlands

[6] 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)

[7] De Clercq, O., Reynaert, M.: SoNaR Acquisition Manual version 1.0. Tech. Rep. LT3 10-02, LT3 Research Group – Hogeschool Gent (2010). URL `http://lt3.hogent.be/en/publications/`

[1] D-Coi deliverables in this list are included in the DOC directory of the SoNaR distribution

[8] Desmet, B., Hoste, V.: Named Entity Recognition through Classifier Combination. In: Computational Linguistics in the Netherlands 2010: selected papers from the twentieth CLIN meeting (2010)

[9] 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, pp. 837–840. LREC-2004, Lisbon, Portugal (2004)

[10] van Gompel, M.: Folia: Format for linguistic annotation. (2011). URL http://ilk.uvt.nl/folia/folia.pdf

[11] 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, pp. 144–149. LREC-2008, Marrakech, Morocco (2008)

[12] Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. Ph.D. thesis, Antwerp University (2005)

[13] 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 (2000)

[14] Karttunen, L.: Discourse referents. Syntax and Semantics **7** (1976)

[15] 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, pp. 1385–1392. LREC-2002, Las Palmas, Spain (2002)

[16] 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) (2007)

[17] Oostdijk, N., Reynaert, M., Hoste, V. and Schuurman, I.: The construction of a 500-million-word reference corpus of contemporary written Dutch. In: Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme. Chapter 13. Springer Verlag. (2013)

[18] Oostdijk, N.: The Spoken Dutch Corpus. Outline and first evaluation. In: Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 887–894. LREC-2000, Athens, Greece (2000)

[19] 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, pp. 1206–1211. LREC-2006, Genoa, Italy (2006)

[20] Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Linguistics Journal **31**(1) (2005)

[21] Reynaert, M.: Character confusion versus focus word-based correction of spelling and OCR variants in corpora. International Journal on Document Analysis and Recognition pp. 1–15 (2010). URL `http://dx.doi.org/10.1007/s10032-010-0133-510.1007/s10032-010-0133-5`

[22] Reynaert, M., Schuurman, I., Hoste, V., Oostdijk, N., van Gompel, M.: Beyond SoNaR: towards the facilitation of large corpus building efforts. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA) (2012)

[23] Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines, version 1.2.1. (2006). URL `http://timeml.org/site/publications/specs.html`

[24] Schuurman, I.: Spatiotemporal Annotation on Top of an Existing Treebank. In: K. De Smedt, J. Hajic, S. Kuebler (eds.) Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, pp. 151–162. Bergen, Norway (2007)

[25] 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 (2007)

[26] 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, Morocco (2008)

[27] 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, pp. 101–112. LINC-2003, Budapest, Hungary (2003)

[28] 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, Malta (2010)

[29] Schuurman, I., Vandeghinste, V.: Spatiotemporal annotation: interaction between standards and other formats. IEEE-ICSC Workshop on Semantic Annotation for Computational Linguistic Resources, Palo Alto, California, USA (2011)

[30] SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language. Version 2.0 MITRE (2007).

[31] 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, pp. 155–158. Taipei, Taiwan (2002)

[32] Trapman, J., Monachesi, P.: Manual for semantic annotation in D-Coi. Technical report, Utrecht University (2006)

[33] Van Eynde, F.: Part of speech tagging en lemmatisering. Protocol voor annotatoren in D-Coi. Project internal document

[34] 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, pp. 1811–1814. LREC-2006, Genoa, Italy (2006)

[35] Woordenlijst Nederlandse Taal (1995). SDU Uitgevers. The Hague, The Netherlands

[36] Woordenlijst Nederlandse Taal (2005). SDU Uitgevers. The Hague, The Netherlands