User requirements analysis for the design of a reference corpus of written Dutch

Nelleke Oostdijk and Lou Boves

Centre for Language & Speech Technology, Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
n.oostdijk|l.boves@let.ru.nl

Abstract

The Dutch Language Corpus Initiative (D-Coi) project aims to specify the design of a 500-million-word reference corpus of written Dutch, and to put the tools and procedures in place that are needed to actually construct such a corpus. One of the tasks in the project is to conduct a user requirements study that should provide the basis for the eventual design of the 500-million-word reference corpus. The present paper outlines the user requirements analysis and reports the results so far.

1. Introduction

Recent surveys that have taken stock of the availability of basic language resources for the Dutch language have identified the need for a large corpus of written Dutch. Daelemans and Strik (2002) observe that, compared to other languages, Dutch is lagging behind. Therefore, the construction of a multi-purpose 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.

On the surface, all stakeholders agree that a large reference corpus of written Dutch would be invaluable for linguistic research and the development of profitable services that require advanced language technology. However, as soon as one starts making preparations for the collection of the text, and the definition of the minimal set of meta-data and annotation layers, it appears that different purposes may very well translate into very different requirements. A very large, balanced, richly annotated multi-purpose reference corpus is very different from the task-specific corpora that have been built in –for example- DARPA programmes and the European CLEF programme. What is more, while some of the stakeholders (e.g. linguists, application developers and system integrators) may be able to formulate requirements and desires in the terms of their own disciplines and business areas, it is not straightforward to translate these formulations into technical requirements for a reference corpus. This is one of the reasons why the Dutch-Flemish STEVIN programme initiated the Dutch Language Corpus Initiative (D-Coi) project.[1]

The aim of the D-Coi project is to specify the design of a 500-million-word reference corpus of written Dutch, and to put the tools and procedures in place that are needed to actually construct such a corpus.[2] It is envisaged that the reference corpus will contribute to the creation of a Dutch HLT infrastructure that is no longer in a position where it is lagging behind, but instead is at the forefront internationally. One of the tasks in the 14-month D-Coi project, therefore, is to conduct a user requirements study that should provide the basis for the eventual design of the reference corpus. The present paper outlines the user requirements analysis and the first results obtained.

2. User requirement study

Although there are as yet no examples of the type of reference corpus that STEVIN is aiming at, it is, of course, possible to derive boundary conditions from experience with existing corpora and the major trends in the development of linguistics and language technology.[3] Thus, a modern reference corpus should not only sample texts from conventional media such as books and newspapers, but also from electronic media, such as web pages, chat boxes, e-mail, etc. It is evident that inclusion of texts from these sources will pose new problems related to IPR, and that they will require the development of novel tools for the detection and annotation of typos, non-words, and similar phenomena that are less important in well-edited texts from the conventional printed media. Another area where the design of a reference corpus of written Dutch can profit from previous experience is the impact that the British National Corpus has had on courses of English as a second and foreign language. However, it is less clear how these boundary conditions and trends translate into the specification of a reference corpus that will meet with general appreciation in several independent communities. Therefore, we decided to follow several parallel paths to complete the user requirements analysis.

First, we defined several different target populations, viz. academic and industrial research in language technology, academic linguists who are studying Dutch as their object language, system integrators and developers of applications relying on language technology, and finally prospective users of products and services that we think would profit from the application of language technology. Next, we developed a web questionnaire in order to elicit from potential users and other interested parties, opinions about several aspects relating to corpus design and annotation. The results were then used in the specification of a provisional design which will be discussed in focus groups that will be conducted shortly and which will involve the different target populations mentioned above. The focus group discussions will also be used to address any issues that remained after the information obtained through the questionnaire had been processed. The insights gained in the process will be cast into a final, detailed specification of the design of the corpus and the annotations to be associated with it.

[1] See D’Halleweyn et al. (2006).

[2] More information about the project can be found on http://lands.let.ru.nl/projects/d-coi.

[3] The American National Corpus is probably closest to what is envisaged for the Dutch reference corpus as it also includes data from electronic media.

3. Questionnaire

3.1. Aims and set up

At the start of the D-Coi project, a website was created that should inform people about the aims of the project and the progress that was being made. For the user requirements study we decided to make use of the website by putting a questionnaire online and invite visitors to the site (interested parties and stakeholders alike) to fill in the form.[4] The questionnaire was set up linearly: questions simply followed one after the other and could be answered – or skipped – independently. There was no obligation to provide an answer to all questions.

In order to promote the project and draw attention to the user requirements study, a presentation was held at the yearly CLIN conference which is attended specifically by computational linguists and language and speech technologists from the Dutch speaking language area. The follow up was done in the form of a mailing through the NTU HLT Newsletter which reaches most people in academia and industry active in the field of Dutch HLT. Several linguists in departments of Dutch, Germanic languages, business communication, and linguistics were approached separately by e-mail.

The first draft (on paper) of the questionnaire was given to two colleagues for feedback on the presentation and nature of the questions. The second, revised version (still on paper) was then used with a small number of potential users who were asked to actually fill in the questionnaire. While the latter group was explicitly told to make suggestions that would improve the questionnaire, no such feedback was given. The questionnaire was then cast into a web form and placed online.

3.2. Aspects targeted

The questionnaire addresses several aspects that relate to corpus design and annotation, as well to the use to which the corpus data is going to be put. More specifically, questions are directed at three areas:

1. who are the respondents: Dutch or Flemish; from academia or industry; active in research and/or teaching; linguists, language and/or speech technologists, etc.; acquainted with the use of corpus data or not?
2. which corpora are currently being used and for what purpose; what is missing?
3. what specific ideas/wishes are there with regard to the design and annotation of a 500-million-word reference corpus of written Dutch? More specifically,
    - what texts ought be included: fiction and/or non-fiction, from what time period and what media, which text types, genres and topics should be represented and by what amount of text, at what audience should texts be targeted, etc.
    - which parts of a text should be included: table of contents, illustrations, captions, running titles, notes, etc.?
    - which text units should one be able to address (chapters, sections, paragraphs, lists, sentences, words)?
    - what annotations are needed?
    - what metadata are required?

It is obvious that the questionnaire emphasized data much more than tools. At the end of the questionnaire, people could volunteer any information they wanted to share.

4. Results

The fact that people were free to answer or skip a question explains why not in all cases an answer was obtained. In the discussion of the results below, therefore, the number of respondents for each question will be indicated separately.

The results as reported in this paper are based on the questionnaires completed by the end of February 2006. They can also be found on the project's website. However, as the questionnaire will remain online for the duration of the D-Coi project, we intend to publish updates of the results at regular intervals.[5]

4.1. Respondents

The questionnaire was completed by 36 respondents. While it is impossible to know how many people can be expected to give an informed opinion about matters relating to corpus design and therefore also what can be considered to be a reasonable number of respondents, we can say that where we aimed to elicit the opinions held by people representing various user groups in both Flanders and the Netherlands, we were quite successful:[6]

- 29 respondents were Dutch and 7 were Flemish
- 30 were from academia vs 6 from industry
- 8 respondents were professors, 16 senior researchers, 7 junior researchers and 5 other
- 18 people were involved in teaching and 34 in research
- 17 were active in the field of linguistics (incl. descriptive linguistics, sociolinguistics, and psycholinguistics), 13 in language- and/or speech technology, 2 in translation studies, while the remaining 4 are concerned with lexicography, communication studies, *e*Humanities and logic
- the range of topics addressed by the respondents is wide and highly varied; topics include spelling, morphology, syntax, semantics, discourse, conversation analysis, L2 acquisition, IR, document classification, machine translation, sociolinguistic variation, machine learning, authorship attribution, data mining and shallow parsing

The fact that the majority of the respondents are Dutch academics reflects the difficulty in involving industrial in medium-term activities, as well as the bigger number of persons active in the field in the Netherlands.

[4] The questionnaire (in Dutch) can be found at http://lands.let.ru.nl/projects/d-coi under 'user requirements study.

[5] See http://lands.let.ru.nl/projects/d-coi.

[6] By comparison: the survey that Bouma and Schuurman (1998) conducted investigating the availability of and need for resources for Dutch was based on information obtained from 40 people, including people from the Dutch Research Foundation, government policy makers and the Dutch Institute for Lexicology. With 32 people interviews were conducted. The group of interviewees then and the respondents of our questionnaire overlap by three people.

4.2. Use of corpus data so far

In the questionnaire a number of questions related to the use of corpus data in research and teaching. From the answers that were obtained, the following picture emerged: a relative large number (27) of respondents use corpus data in their research. Those who are involved in research and do not use corpus data, all except one intend to do so in the future. Rather different is the situation with regard teaching: eight of the 18 respondents that are involved in teaching find that corpus data are irrelevant to the subject matter they teach.

Since we wanted to know what corpora are already being used and also what their strengths and weaknesses are as experienced by users, the respondents were presented with a list (cf. Table 1) comprising most of the corpora more or less publicly available for Dutch and then asked to indicate which corpora they used in their research and/or their teaching. In case respondents used other corpora or data collections, they could list these separately.[7] An overview of the Dutch corpora that respondents use in research (R) and teaching (T) is given in Table 1.

Corpus	R	T
Corpus Uit den Boogaart	7	5
38 MW Corpus INL	7	2
Parole Corpus	3	3
CLEF data collection	2	–
Twente News Corpus (TwNC)	5	2
Mediargus/KNAC	1	1
CONDIV	5	–
CoGen	2	–
CGN	20	10
Other	18	6

Table 1: Corpora in use in research (R) and teaching (T)

As it turned out, quite many respondents said that they used private collections or data made available to them by the content owners for use within a specific project. Of course where highly specific data are required (as e.g. in the case of a study of L2 acquisition by Turkish and Moroccan learners of Dutch), this can be expected. However, as respondents indicated, the creation of private collections has often been triggered by the fact that they were not satisfied by what available corpora had to offer. 17 of the respondents who use corpora for research purposes found that the available corpora fall short when it comes to (a) the amount of data, (b) the quality of the data, and/or (c) the (quality or availability of) annotations.[8] Similar findings hold with regard to the use of corpora in teaching.

Information need

In order to elicit information as to what it is that people using corpora are looking for, respondents were presented a list of possibilities (cf. Table 2). In the case of derived information, i.e. frequency and distribution information, respondents were asked to specify their information need. The majority of respondents here indicated the need for frequency information about basically everything: words (types, tokens), lemmata, compounds, multi-word expressions, word class membership, spelling errors, etc. Some also explicitly indicated the need for sub-word level information (incl. morphemes, syllables, letters but also stress). Typically, technologists brought to bear the need for *n*-gram information.

What information	# times mentioned
Tokens	28
Multi-word expressions	25
Syntactic structures	21
Frequency information	26
Distribution information	19
Other	7

Table 2: Information need (max N=34)[9]

4.3. Design

Nature of the data

Respondents were asked to indicate whether in their opinion the corpus should comprise fiction and/or non-fiction texts. Moreover, in the latter case they should specify which of the types listed (informative, persuasive, instructive texts) should be included. Respondents could also enter non-listed alternatives. The responses are summarized in Table 3.

Fiction/non fiction	# times mentioned
Fiction texts	26
Non-fiction: informative texts	32
Non-fiction: persuasive texts	28
Non-fiction: instructive texts	32
Other	3

Table 3: Fiction and/or non-fiction? (max N=35)

Most respondents agree on the inclusion of non-fiction texts. Remarkable was the fact that some people (typically language and speech technologists) explicitly indicated not to see any use for including fiction. Suggestions were made to include also private texts, texts produced by L2 learners of Dutch and texts produced by bilingual writers.

Time period

In response to the question from what time period texts should be collected, the respondents appear to be divided. While the majority of respondents indicated that they prefer texts from 1980 onwards, there are also quite a number of people who would like to see texts included dating from 1950 onwards. There appears to be a tendency for technologists to prefer more recent material as opposed to linguists who, on the whole, are in favour of including texts from a wider time span.

[7] These have been included under ‘other’ in Table 1.

[8] ‘Quality of the data’ generally refers to the fact that the corpora often include data from a very limited number of text types.

[9] *N* or *max N* indicates the number of respondents that answered the question.

Time period	Fiction	Non-fiction
1950 - present	8	7
1960 - present	2	2
1970 - present	0	4
1980 - present	5	7
1990 - present	10	11
other	10	4

Table 4: Time period from which to collect texts (N=35)

Note that in the case of fiction, the number of 'other' is accounted for by respondents (with the exception of one) who do not see the need to include fiction. In the case of non-fiction, three respondents argued the need for including diachronic data.[10] One respondent explicitly suggested focussing on the same period represented in the spoken Dutch Corpus (CGN).

Media

The questionnaire explicitly addressed the question whether the corpus should include texts from both conventional and newer, electronic media. Inclusion of texts from conventional media appears beyond dispute. A vast majority of respondents (28 out of 34) were also in favour of including texts from newer media. Interestingly, here some respondents make of point of stating that it is not just any text type available though the newer media that should be included.[11] Other remarks offered by the respondents suggested the inclusion of Dutch learner data, essays and other writing products of students and pupils and elicited data.

Text types

In order to elicit ideas of what text types should be incorporated, respondents were presented with a list of various text types (cf. Table 5). They were asked to indicate which of these they deemed desirable and also what they thought were minimum quantities that were required. Only 13 respondents gave a full answer addressing both aspects.[12] Nine respondents failed to provide an answer, some of them stating that they found the question too difficult and had no idea whatsoever. Another 14 respondents only completed the question in so far as it related to the text types to be included; they were not able to specify any quantities. Table 5 lists the text types and the number of respondents favouring their inclusion.

As the figures in Table 5 show, there is little agreement as to what text types from the newer media should be included. There appears to be a tendency for linguists (as opposed to technologists) to be in favour of including highly informal text types such as e-mail, msn, and sms.

With regard to the minimum number of words that are required for a given text type, opinions were highly divided. Whereas technologists suggested amounts in the ranges of several up to 50 million words, linguists were thinking of 50,000-500,000 words.

Text type	# times mentioned
Printed books: fiction	18
Printed books: non-fiction	22
Printed magazines	20
Printed newspapers	24
Printed folders and brochures	16
Printed reports	17
Printed summaries	14
Websites	18
Discussion lists	10
E-mail messages	17
E-books	11
E-magazines	11
MSN texts	16
Subtitles (television)	12
Autocues	9
Other	3

Table 5: Text types to be included (max N=27)

Sample sizes

The total size of the reference corpus envisaged is 500 million words. This would allow for a fair number of full texts to be included. However, if we are looking for balance, it might well be that the inclusion of only full texts is not the best option. We therefore asked respondents which they thought was the better option: (a) always include full texts (regardless its length), or (b) in the case of shorter texts, always include the full text and in the case of longer texts select a sample (comprising e.g. the first n chapters, an x number of words). We also invited other suggestions. Unfortunately, due to a technical error no responses were logged for this question.

Audience targeted

There were two questions in the questionnaire that related to the audience targeted. Thus, respondents were asked to indicate what proportion of texts should be directed at what age groups (adults, teenagers, children). A second question asked whether texts to be included should be directed at (a) the general public, (b) peers and colleagues, (c) students and pupils, (d) clients, and (e) some other audience (to be specified).

For the first question, 28 respondents provided an answer. Five of these opt for only texts directed at adults, excluding texts directed at teenagers and children. Thirteen respondents opt for over 70% of adult texts, eight for 50-60%, and two for 33 and 25% resp. Eight respondents are in favour of excluding texts directed at teenagers, while nine do so for texts directed at children. Proportions suggested for texts directed at teenagers range between 5 and 25% (with one exception: 33%). The same holds for texts directed at children.

The second question was answered by 33 respondents. They appeared to be divided as to what readership the texts should be targeted at. While some respondents appear to be fervently in favour of including only texts directed at a general public, others are interested in texts for specific groups of readers only. Preferences cannot be generalized over either linguists or technologists as a group.

[10] One respondent suggested to include even older texts.

[11] The issue came up again as the respondents were asked to indicate what text types should be included. See under text types.

[12] Nine of these were language- and/or speech technologists, 4 were linguists.

Readership	# times mentioned
General public	20
Peers, colleagues	14
Students, pupils	14
Clients	14
Other[13]	1

Table 6: Targeted readership (max N=33)

Genres and topics

Respondents proved very reluctant in putting forward their ideas with regard to the specific genres or topics they thought the corpus should feature. There were some suggestions to include texts from the medical, legal and science domains, and also to include professional language. On the other hand, some respondents also made it clear that they would rather not include texts containing a lot of jargon.

Text elements

We presented respondents with a list of text elements (incl. titles of chapters and (sub)sections, preface, table of contents, index(es), bibliography, illustrations, notes, captions and running titles) and asked them to indicate which of these should be included. Most of the 29 respondents who gave an answer were in favour of including the titles of chapters and sections and the preface of text. Only 14 indicated that in their opinion notes and captions should also be included. None of the other elements were deemed essential.

Units for selection

There was also a question relating to the units in a text that people should be able to search for. A vast majority of the respondents who answered the question considered only the accessibility of words, sentences and paragraphs essential (cf. Table 7).[14]

Unit	# times mentioned
Chapters and (sub)sections	20
Paragraphs	29
(Un)ordered lists	16
Sentences	30
Words	34

Table 7: Units for selection (max N=35)

Annotations

With regard to the types of annotation one wanted to have available with the corpus, a remarkably large number of respondents indicated the need for syntactic annotation. Semantic annotation was the one single type of annotation named under ‘other’. Here, however, it should be noted that the answers also indicate little consensus as to the exact nature of this type of annotation.

Annotation	# times mentioned
POS tagging	27
Lemmatisation	23
Decompounding	20
Labelling of multi-word expressions	26
Syntactic annotation	26
Other	7

Table 8: Annotations (max N=30)

Metadata

All 36 respondents gave their opinion as to what metadata they thought should be available with the data. Somewhat surprisingly, not everyone considered metadata relating to the author (name, sex, age, etc.) and the publication (title, publisher, publication data, original language, etc.) relevant. Respondents made specific suggestions about including also information about language proficiency levels and the language background of learners. Several respondents indicated that they would like to see the design of the corpus reflected in the metadata, such that texts belonging to one and the same component in the corpus can easily be selected.

Metadata	# times mentioned
Author-related	31
Publication related	31
Translator related (where appropriate)	25
No. of words for a given text	23
Info about available annotations	24
Other	9

Table 9: Metadata (max N=36)

4.4. Respondents’ comments

At the end of the questionnaire, respondents were asked to enter any remaining comments and/or suggestions. The one comment recorded most often referred to the degree of difficulty of the subject matter and the questions asked. This in fact confirms what was already apparent from the fact that only very few questions were answered by all respondents. One respondent stated that “many questions had better be put to an expert”. Some respondents used this opportunity to point out that only original Dutch texts should be included, while others would argue in favour of also including Dutch translations.

5. Provisional corpus design

On the basis of information available from other corpus initiatives and knowledge of existing Dutch resources, combined with the input obtained through the questionnaire, a provisional corpus design was made. The design presently under consideration is shown in Figure 1.

The design aims at a reference corpus of contemporary standard written Dutch as encountered in texts (i.e. stretches of running discourse) originating from the Dutch speaking language area in Flanders and the Netherlands as well as Dutch translations published in and targeted at this area. The corpus will include learner and native speaker language and the language of (professional) translators.

[13] The suggestion made here was to distinguish also texts directed at family, friends, social peer group, etc. (esp. for learner data).

[14] The units were prelisted. Several respondents here noted that it should also be possible to have access to morphological elements, not realizing that this requires annotation rather than mere markup.

6. Follow-up

The provisional design will be presented to a number of focus groups along with the issues that were raised by the questionnaire. We intend to organize meetings with special interest groups, such as teachers and course developers in the field of Dutch as a second language. In collaboration with the Dutch Organization for Language and Speech Technology (NOTaS) we will conduct sessions with system integrators and application developers, in which we will discuss the technical and economical viability of novel HLT products. Similar sessions will be conducted with a group of people who are responsible for process and product innovation in large organizations such as banks, utilities and government agencies.

Written to be read, published, electronic	
Discussion lists	2.5 MW
E-books	5 MW
E-magazines	25 MW
E-mail (spam)	2.5 MW
Newsletters	2.5 MW
Press releases	10 MW
Subtitles	10 MW
Teletext pages	50 MW
Websites	50 MW
Wikipedia	20 MW
Written to be read, published, printed	
Abstracts, summaries	10 MW
Books	75 MW
Brochures	5 MW
Newsletters	2.5 MW
Guides, manuals	5 MW
Legal texts	2.5 MW
Newspapers	50 MW
Periodicals, magazines	10 MW
Policy documents	5 MW
Proceedings	10 MW
Reports	5 MW
Surveys	2.5 MW
Theses	2.5 MW
Written to be read, unpublished, electronic	
Chats	25 MW
E-mail (non-spam)	50 MW
Minutes	10 MW
Sms	5 MW
Written assignments	10 MW
Written to be read, unpublished, printed	
Theses	10 MW
Written to be read, unpublished, typed	
Minutes	10 MW
Written assignments	10 MW
Written to be spoken, unpublished, electronic	
Autocues	2.5 MW
Written to be spoken, unpublished, typed	
News scripts	2.5 MW
Texts for the visually impaired	2.5 MW

Figure 1: Provisional corpus design

The results of the sessions with the groups mentioned above will be cast into a final, detailed specification of the design of the corpus and the annotations to be associated with it. Preparations for the focus groups are presently under way.

7. Conclusion

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

As experiences with the questionnaire show, it is not easy to elicit views on the composition of a corpus yet to be designed. Apparently, for the user in general the distance between actual use of the corpus on the one hand, and its design on the other hand, is rather, if not too big. Where the designer is concerned with and distinguishes between requirements with regard to more specifically the raw data, mark-up, annotations, and metadata, much of this appears to escape the user. Yet, to conclude on a more positive note, the questionnaire has proven to be extremely helpful in identifying the issues that require further attention. We expect that discussions in focus groups can be conducted in such a fashion that conclusive arguments for or against a particular line of action will emerge.

8. Acknowledgement

The D-Coi project is funded by the NTU/STEVIN programme (http://www.taalunieversum.org/stevin) under grant number STE04008. Thanks are due to the respondents who completed our questionnaire and thus provided us with invaluable information.

9. References

Bouma, G. and I. Schuurman (1998). *De positie van het Nederlands in Taal- en Spraaktechnologie.* http://odur.let.rug.nl/~gosse/taalunie/webrapport/

Daelemans, W. and H. Strik (2002) *Het Nederlands in de taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen.* Nederlandse Taalunie.

D'Halleweyn, E., J. Odijk, L. Teunissen, and C. Cucchiarini (2006). The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources. In *Proceedings LREC 2006.*

Grondelaers, S., K. Deygers, H. Van Aken, V. Van Den Heerde, D. Speelman (2000). Het CONDIV-corpus geschreven Nederlands. In *Nederlandse Taalkunde*, nummer 2, 2000 and *DigiTaal*, No. 11 (Available from http://www.niederlandistik.fu-berlin.de/digitaal/).

Oostdijk, N. (2000). The Spoken Dutch Corpus. Outline and first evaluation. In *Proceedings of LREC'00.* Athens, Greece. Vol. 2.: pp. 887-894.