TwNC: a Multifaceted Dutch News Corpus

Roeland Ordelman, Franciska de Jong, Arjan van Hessen, Hendri Hondorp.

University of Twente (UT)

Department of Electrical Engineering, Mathematics and Computer Science

Human Media Interaction Group

P.O. Box 217, 7500 AE Enschede

The Netherlands

http://hmi.ewi.utwente.nl

{fdejong, ordelman, hessen}@ewi.utwente.nl

Introduction

This contribution describes the Twente News Corpus (TwNC), a multifaceted corpus for Dutch that is being deployed in a number of NLP research projects among which tracks within the Dutch national research programme MultimediaN, the NWO programme CATCH, and the Dutch-Flemish programme STEVIN. The development of the corpus started in 1998 within a predecessor project DRUID and has currently a size of 530M words. The text part has been built from texts of four different sources: Dutch national newspapers, television subtitles, teleprompter (auto-cues) files, and both manually and automatically generated broadcast news transcripts along with the broadcast news audio. TwNC plays a crucial role in the development and evaluation of a wide range of tools and applications for the domain of multimedia indexing, such as large vocabulary speech recognition, cross-media indexing, cross-language information retrieval etc. Part of the corpus was fed into the Dutch written text corpus in the context of the Dutch-Belgian STEVIN project D-COI that was completed in 2007. The sections below will describe the rationale that was the starting point for the corpus development; it will outline the cross-media linking approach adopted within MultimediaN, and finally provide some facts and figures about the corpus.

Text corpora and the development of Semantic Access to Multimedia via Natural Language

To tackle the challenge handed in by the ever increasing volumes of multimedia content being created and stored on the one hand, and the promising steps made in multimedia analysis on the other hand, it is tempting to put the focus on the advancement of image analysis and its integration with recently gained insights from relevant domains such as knowledge extraction

and semantic web technology. Still, the rationale within programmes such as MultimediaN is that it would be grossly erroneous to ignore the available insights in the successful role that can be played by the modality for which very matured analysis frameworks exist: natural language.

As is widely acknowledged, the exploitation of linguistic content in multimedia archives can boost the accessibility of multimedia archives enormously. Already in 1995, Brown et al[2] demonstrated the use of subtitling information for retrieval of broadcast news videos, and in the context of TRECVID a common feature of the best performing video retrieval systems is that they exploit speech transcripts. Of course the added value of linguistic data is limited to video data containing textual and/or spoken content, or to video content with links to related textual documents, e.g., subtitles, manually generated transcripts, etc. But, where available, linguistic content can play a crucial role bridging the semantic gap between media features and user needs.

Depending on the resources available within an organization that administers a media collection, the amount of detail of the metadata and their characteristics may vary. Large national audiovisual institutions annotate at least descriptive metadata: titles, dates and short content descriptions. However, many multimedia archiving institutes often do not have the resources to apply even some basic form of archiving. In order to still allow the conceptual querying of video content, collateral textual resources that are closely related to the collection items can be exploited. They can be either available because they play a role in the production or broadcast process, or they can be generated via speech recognition. Collateral data can be exploited to (i) produce highly accurate, automatic indexes in an affordable way, (ii) tune speech and language processing tools to the focus domain, and (iii) enhance presentation of video retrieval search results by adding extra layers of information.

A well known example of such a collateral textual resource is subtitling information for the hearing-impaired (e.g., CEEFAX pages 888 in the UK) that is available for the majority of contemporary broadcast items, and in any case, for news programs. Subtitles contain a nearly complete transcription of the words spoken in video items and can be easily linked to the video by using the time-stamps that come with the subtitles. Textual sources that can play a similar role are teleprompter files: the texts read from screen by an anchor person (also referred to as auto-cues). Also outside the broadcast sector, collateral materials that somehow match the speech in a collection can be found. A collection of recorded lectures may have presenter notes attached to it, speeches may be accompanied with the written text version, and for meeting recordings there may be minutes available, or at least an agenda.

In absence of (or lack of access to) such error-free texts, there is always the possibility to use automatic speech recognition (ASR) for the generation of transcripts. Relatively limitedly referenced, however, is the exploitation potential for textual content to complement speech transcripts. In all the examples mentioned above the time stamps in the sources are crucial for the creation of a textual index into video. In collateral text sources, the available time-labels are not always fully reliable and outside the news broadcast domain they will often be absent. The resynchronization or labelling of the text with time-codes is called the ‘alignment’ of text and speech, a well-known procedure used frequently in ASR, for example, when training acoustic models. The alignment of collateral data holds for surprisingly low text-speech correlation levels, especially when some additional trickery is applied.

Recent years have shown that large vocabulary speech recognition can successfully be deployed for creating multimedia annotations allowing the conceptual querying of video content. When the collateral data only correlates with the speech on the topic level, full-blown speech recognition must be called in, using the collateral data as a strong prior ('informed speech recognition') or source for extensive domain tuning via the language model. Alternatively, the collateral data can be used for relevance feedback during search. Also the out-of-vocabulary rate could be decreased: if a (non-perfect) ASR transcript is used as the basis for a search of related text, and the terms referring to named entities in the most similar texts are fed into the language models, a second run of the ASR could yield improved recognition results.

Ideally one would not only synchronize audiovisual material with content that approximates the speech in the data, but take even one step further and exploit any accessible text including open source titles and proprietary data (e.g., trusted web pages and newspaper articles). In the context of meetings for example, usually an agenda, documents on agenda topics and CVs of meeting participants can be obtained and linked to the media repository.

Finally, there is of course also the possibility to use an audio fragment as a query for textual documents. Via a transcription of an audio query, related text can be identified. An obvious application domain for this option is, again, news. But it works also in other domains than news, e.g., oral history archives, meeting or lecture recordings, audio blogs, digital storytelling, etc.

The structure of the Twente News Corpus

The original goal for starting the development of the Twente News Corpus (TwNC) was to collect data for the training of language models and acoustic models to be incorporated into a system for large vocabulary speech recognition for Dutch to be deployed in the broadcast news domain and, also, as a baseline system in other domains that lack large amounts of example data (e.g., cultural heritage data as we encounter in the Dutch CHoral project). The focus on news was given in by the size of the datasets available for this domain, and by the focus on news as target at many other research groups. News is a target domain for corpus development, for search applications and for speech technology.

Several requirements come from this type of deployment for a text corpus. They pertain to formatting, encoding, size and balancing, for example. TwNC text data has been formatted as XML and the encoding chosen is utf-8. Balancing is reached by selecting four different source types: newspaper text, autocue files (teleprompter text), subtitling files and manually generated transcripts. The current corpus size is approximately 500 million words of text and about 800 files of broadcast news audio. In the remainder of this section we will describe the types in more detail.

Newspaper data

One of the largest publishers in the Dutch language region, PCM publishers, have donated content on a daily basis (via ftp) from six national newspapers. UT was given access to two years ('94-'95) of content from *NRC Handelsblad* and *Trouw*. Since 1999 also content from four other

newspaper titles is made available: *Algemeen Dagblad*, *Volkskrant*, *Parool* and *NrcNext*. There are even a few years for which content is available from magazines such as *Vrij Nederland* and *HP De Tijd*, and soon also *Groene Amsterdammer*. In Table 1 and Table 2 the statistics of the newspaper data are listed. Daily newspaper feed is not just helping to enlarge the corpus, it also facilitates the updating of the broadcast news vocabulary, or the daily production of word occurrence statistics and predictions such as illustrated in Figure I. A number of research groups have explored parts of the newspaper corpus for tasks such as measuring lexical variation, paraphrase identification learning, and extracting hypernym-hyponym relations.

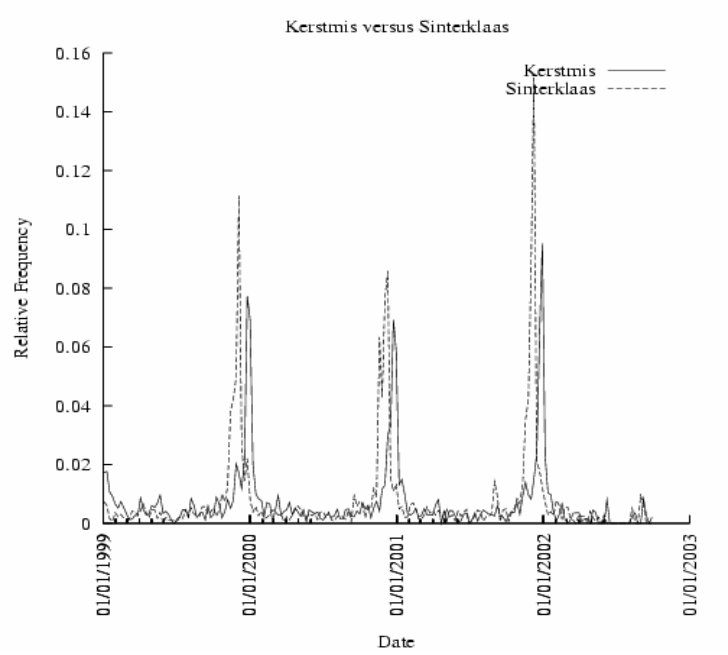


Figure 1 Word occurrence statistics of the words Christmas and Santa-Claus from 1999-2003

Year	Mwords
1994	34

1995	33
1999	63
2000	82
2001	70
2002	70
2003	34
2004	40
2005	36
2006	41
2007	37
2008	7
Total	**547**

Table 1 Number of newspaper words (in millions) per year

Newspaper/ Magazine	**NPs**	**Mwords**
NRC Handelsblad	2913	156
Volkskrant	2392	137
Algemeen Dagblad	2375	102
Trouw	2020	85
Parool	1475	58
NrcNext	169	5
Dordtsch Dagblad	117	2
HP De Tijd	21	1

Vrij Nederland	18	1
Total	**11500**	**547**

Table 2 Number of newspapers and words per newspaper/magazine since 1994

Subtitles

Since 1998 we have been capturing subtitling for the hearing impaired of broadcast news shows that are normally projected on the television screen in the Netherlands via teletext page 888 (CEEFAX pages 888 in the UK) but can also be accessed directly using a TV-card.

Due to a minimum of available space for subtitles on a screen, the number of words in the subtitles is cut down drastically compared to what was actually said. Although phrases are often mixed up completely in an attempt to say the same with less and often other words, subtitles provide an excellent information source for automatic indexing. Moreover, within the Dutch broadcast news autocue files, subtitle topics are marked which is very interesting from a low cost indexing perspective. Finally, the subtitles have time information attached to them referring to the exact time the subtitle was projected on the screen. The delay with respect to the speech differs depending on whether the subtitles are generated live or not. The broadcast news retrieval demonstrator[3] that is running at the University of Twente shows how both speech recognition and subtitle information can be deployed for indexing a news show.

The captured teletext subtitles have been converted to a suitable XML format that includes the topic boundaries and time information. In total, the Twente News Corpus has 2.3 M words of broadcast news subtitles. For a subset of this collection also the audio of the broadcast is available.

Teleprompter files

The third type of news related text data we collected consists of *teleprompter* files also referred to as *autocues*: the texts the newsreaders read from the teleprompter. The autocues have been kindly provided between 1998 and 2005 by the Dutch National Broadcast Foundation (NOS), the producer of the 8 o'clock news show (8 Uur Journaal). The autocues are almost an exact representation of the speech from the newsreaders but lack of course spontaneous utterances and live commentaries ('pseudo-live' commentaries are often included). There is topic information and (relative) time information available in the files. The RTF format autocues were converted to XML and have a total word count of 3.3 M words.

Transcripts and audio

For the purpose of training acoustic models for a broadcast news speech recognition system, 26 broadcast news shows were recorded and manually annotated on the word level. From 2005 onward, both audio and subtitles from the 8 o'clock news shows that were recorded and indexed in our broadcast news retrieval demonstrator were preserved. These data pairs (currently some 800) can then for example be used for the partly unsupervised training of acoustic models. With a non optimised routine we automatically extracted about 60 hours of training data consisting of a lot of small audio segments with aligned transcripts (sequence of 3 words minimum) from 279 news shows (2006). UT is currently also processing other years and intends to make the acoustic models that are trained with the data available (see also 'Tools' below).

	with transcripts	**with subtitles**	**with autocues**
Audio (828)	26	801	30

Availability Conditions

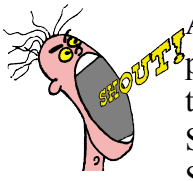
The newspaper content is made available for use by researchers under the condition that they do not publish any summaries, analyses or interpretations of the linguistic characteristics that can lead to extraction of reconstruction of the original content. UT is allowed to redistribute portions of the data under strict licence agreements. Currently, access to the 1999-2002 data can be licensed to individual research groups. The 1994-1995 data was redistributed among the participants of the evaluation campaign within CLEF (Cross-Language Evaluation Forum). Recently, the 1999-2004 data was made available to participants of the Dutch STEVIN speech recognition benchmark evaluation N-BEST, specifically for purposes of language model research.

Table 1 Overview Twente News Corpus

Source	**amount**
Newspapers	547Mw
BN Subtitles	2.3Mw
BN Autocues	3.3Mw
BN Audio	828
BN Transcripts	26

BN Alignments	60+ hours

Tools



Along with the corpus itself UT can provide tools developed for text normalisation purposes and speech recognition: the UT Text Normalisation Toolkit (UT-TNT), that was partly developed in the MultimediaN project and the STEVIN project SPRAAK, and the SHoUT speech recognition toolkit developed in MultimediaN. SHoUT Acoustic models that are trained using the automatic alignment procedure will be made available at a later stage.

Links

1) Twente News Corpus: http://hmi.ewi.utwente.nl/twnc

2) MultimediaN: http://www.multimedian.nl

3) NWO programme CATCH: http://www.nwo.nl/catch (in Dutch only)

http://hmi.ewi.utwente.nl/showcases/

4) STEVIN: http://www.taaluniversum.org/stevin (in Dutch only)

5) Project DRUID: http://hmi.ewi.utwente.nl/Projects/druid.html

6) STEVIN projects D-COI & SPRAAK: http://hmi.ewi.utwente.nl/project/STEVIN

7) Choral Project: http://hmi.ewi.utwente.nl/choral

8) Cross-Language Evaluation Forum (CLEF): http://www.clef-campaign.org/

9) Shout Speech Recognition Toolkit: http://www.vf.utwente.nl/~huijbreg/shout/index.html

10) TRECVID: http://www-nlpir.nist.gov/projects/trecvid/

1 Relevant links are listed at the end of this article.

2 M. G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S. J. Young. Automatic Content-based Retrieval of Broadcast News. In Proceedings of the third ACM international conference on Multimedia, pages 35–43, San Francisco, November 1995. ACM Press.

3 http://hmi.ewi.utwente.nl/showcases/broadcast-news-demo