Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus

Jeska Buhmann [(1)], Johanneke Caspers [(2)], Vincent J. van Heuven [(2)], Heleen Hoekstra [(3)], Jean-Pierre Martens [(1)], Marc Swerts [(4)]

[(1)] Universiteit Gent,
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
{Buhman, Martens}@elis.rug.ac.be

[(2)] Phonetics Laboratory, Universiteit Leiden,
Cleverinaplaats 1, PO Box 9515, 2300 RA Leiden, The Netherlands
{J.Caspers, V.J.J.P.van.Heuven}@let.LeidenUniv.nl

[(3)] UiL-OTS, Universiteit Utrecht,
Trans 10, 3512 JK Utrecht, The Netherlands
H.Hoekstra@let.uu.nl

[(4)] CNTS, Universitaire Instelling Antwerpen
Universiteitsplein 1, B-2610 Wilrijk, Belgium
swerts@ipo.tue.nl

Abstract

This paper first describes the aims of the prosodic annotation for (part of) the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN), and the procedures that are currently being developed to produce the annotation. It further reports on a pilot study that was run to estimate the costs and the attainable quality (in terms of inter-transcriber consistency) of the envisaged annotation. It is our claim that high-quality prosodic annotation (of prominence, prosodic breaks, and unusual segmental lengthening) can be obtained by non-experts, provided these are given a strict, written protocol and a short period of supervision and feedback.

1. Introduction

Understanding the prosodic mechanisms dominating the spoken communication between humans is of great importance for the further development of human-machine dialog systems. It is generally acknowledged, though, that in order to make real progress in this area, one needs large, prosodically labeled corpora. Since there are currently no such corpora available for Dutch, it was decided that a subset of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) will be prosodically annotated.

The Spoken Dutch Corpus (CGN) is going to be a large compilation (about 10 million words or 1,000 hours of speech) of Dutch as it is spoken in The Netherlands and in Flanders (in a 2:1 proportion). It is being developed for a multi-disciplinary user group, and it is going to contain speech from various socio-situational settings. All speech will be orthographically transcribed, lemmatised and enriched with part-of-speech information. For one million words, more detailed information will be provided, such as a broad phonetic transcription, a manually verified word segmentation and a syntactic analysis (Oostdijk et al., 2002). A quarter of the one million words (250,000), or 25 hours of speech, will also receive a prosodic annotation. This subset will be divided into two equally large parts: a Dutch and a Flemish part.

This paper first describes the aims of the prosodic annotation and the procedures that are being developed to produce it, it then reports on the results of a pilot study that was run to estimate the costs and the attainable quality (in terms of inter-transcriber consistency) of the envisaged annotation.

In the process of defining the aims of the prosodic annotation, potential users with expertise in the domain were consulted. After some discussion, it soon became transparent that a fine-grained labeling like ToDI (Gussenhoven et al., 1999) would be impossible to achieve within the budgetary constraints. There was a clear consensus for preferring a large corpus with less detailed annotations over a smaller corpus with more refined annotations. Other arguments against ToDI were that it is too theory dependent, and that it requires well-trained transcribers. Therefore, it was decided to envisage a perceptually-based annotation as in Portele & Heuft (1997) and Grover et al. (1998) instead. The key elements to be labeled are prominence, prosodic boundary strength and (unusual) segmental lengthening.

Given the limited resources (time and money) available, and given the limited availability (and willingness) of experts to perform the task, it became clear that we would have to rely on non-expert transcribers. In view of all this, we decided to give the transcribers the following tasks:

1. Mark syllables which are carrying a clear prominence.
2. Mark important between-word and within-word interruptions of the normal speech stream (henceforth called ‘breaks’) as either weak or strong breaks.

3. Mark unusual lengthening of individual sounds which are not causing prominence.

As there was evidence (e.g. Streefkerk et al., 1997) that even such a simple task was not automatically going to lead to consistent annotations, it was decided to pay attention to the development of a protocol for prosodic annotation, and to run a pilot study in order to assess the validity of the proposed approach before starting any large-scale production of annotations.

The remainder of this paper is organized as follows. Section 2 discusses the preparation of the data, and the rules and procedures that are outlined in the protocol for prosodic annotation. Section 3 describes the goals of the pilot study, the experiments that were carried out in order to achieve these goals, and the results of the pilot study that emerged from the analysis of the annotation data. The paper ends with a discussion and a proposal for the final production of the annotations.

2. Procedures

This section of the paper describes the preparatory stages of the project. The preparations comprised the construction of an efficient on-line working environment (user interface for audio-visual display of waveforms and time-aligned text files) for the transcribers and the compilation of a written instruction protocol that could be studied off-line.

2.1. Data preparation

A multi-layered text file was prepared such that orthographic transcripts of all audio files were automatically synchronized to a waveform display of the signals to be annotated. All prosodic annotations were to be entered by the transcribers in these text files. The text files were initialized with time markers and orthographic transcripts of the stretches of speech that were spoken between these time markers. The time markers delimit stretches of speech that are separated by long pauses (defined as stretches of signal without a transcript).

The initialization of the prosodic annotation files was performed in two steps.

Step 1. So as to favor the perceptual nature of the annotation and to suppress any bias towards putting breaks at syntactic boundaries, all punctuation marks were removed from the orthography.

Step 2. Prosodic annotations would only be performed on files for which a manually verified word segmentation was available. Since this segmentation also identifies clear pauses between words (Martens et al., 2002), it was possible to employ an automatic phrasing system to split up the speech in phrase-like units on the basis of the word segmentation.

The automatic phrasing system was designed in such a way that it produces units that are no longer than 10 seconds, and that are separated by long pauses. The algorithm runs from left to right through the signal. Given a temporary starting time, it searches for the first pause that is longer than 0.5 seconds, and puts a phrase boundary at the start of that pause. However, if a phrase turns out to be longer than 10 seconds, the algorithm backtracks to the longest pause within the most recent 10-second interval, and takes the onset of this pause as the phrase boundary. Once a new phrase boundary is located, the temporary starting time is moved to the end of the pause following that boundary, and the algorithm continues until the end of the file is reached.

It is assumed (and verified on an evaluation corpus) that all the phrase boundaries correspond to perceptually strong breaks, and that they need no verification.

2.2. The user interface

The manual annotation is performed using the Praat tool (Boersma & Weenink, 1996; Boersma & van Heuven, 2001). The transcribers are looking at a computer screen with a display of the signal and its orthographic transcript. The orthographic text (without any punctuation marks) is organized in speaker tiers (one tier per speaker), synchronized with the signal, and presented as a sequence of phrase-like units (having a transcript) and pauses (having no transcript, see figure 1, next page). The phrase boundaries appear as blue (black in figure 1) vertical lines in the orthographic tiers

2.3. The protocol

The protocol starts with describing the aims of the annotation, the properties of the supplied orthographic transcripts, and the basic principles underlying the annotation procedure. Some of the general guidelines are:

1. The prosodic phenomena are marked by special symbols, which are inserted in the orthographic tiers.
2. No changes in the orthography are allowed other than the **insertion** of prosodic symbols,.
3. Between-word break symbols must be surrounded by spaces; within-word break symbols should not.
4. For recordings of multi-party types of discourse, the transcriber should first annotate all the speech of one interactant before turning to the second speaker.

After having outlined these principles, the protocol continues with a description of the four phenomena to be annotated. For each of these, there is a set of rules to guide the transcriber in case of doubt. Let us briefly review the four phenomena:

1. **Strong breaks** (symbol '||') are defined as severe interruptions of the normal flow of speech. They are typically realized as a clear pause or even an inhalation.

   Ex: *he was there || and so was his girl-friend*

2. **Weak breaks** (symbol '|') are defined as weak but still clearly audible interruptions of the speech flow. Although no real pause is observed, it is clear that the words (or parts of a word) straddling the break are not connected the way one would expect them to be in fluent speech. In case of doubt between a strong and a

weak break, the human transcriber is instructed to choose for a weak break.

Ex: *I can tell you | this was un|be|lievable*

3. **Prominent syllables** (symbol '^') are defined as syllables that are emphasized by the speaker, e.g. to make a word important (i.e., to put it in focus). Prominence is typically realized by a pitch movement, often in combination with vowel lengthening and/or an increase of loudness. In case of doubt, the human transcriber is suggested to try and repeat the phrase with and without prominence on the target syllable, and to decide which realization is most similar to the one heard in the speech file. A prominent syllable is marked by putting the orthographic characters corresponding to the phonetic vowel nucleus between prominence symbols.

lengthening. In case of doubt between prominence and segmental lengthening, transcribers were instructed to indicate prominence. As in the case of prominence, lengthening symbols are put around the orthographic characters corresponding to the sound that is being lengthened.

Ex: *no || he's just fift%y% f^ou^r now*

The protocol ends with some suggestions on how to navigate through the files, what to mark first, etc.

3. The pilot study

The pilot study was run in four universities: two in The Netherlands (Leiden, Utrecht) and two in Flanders (Antwerp, Ghent). Four pairs of naive transcribers (one pair per university) were hired for this study.

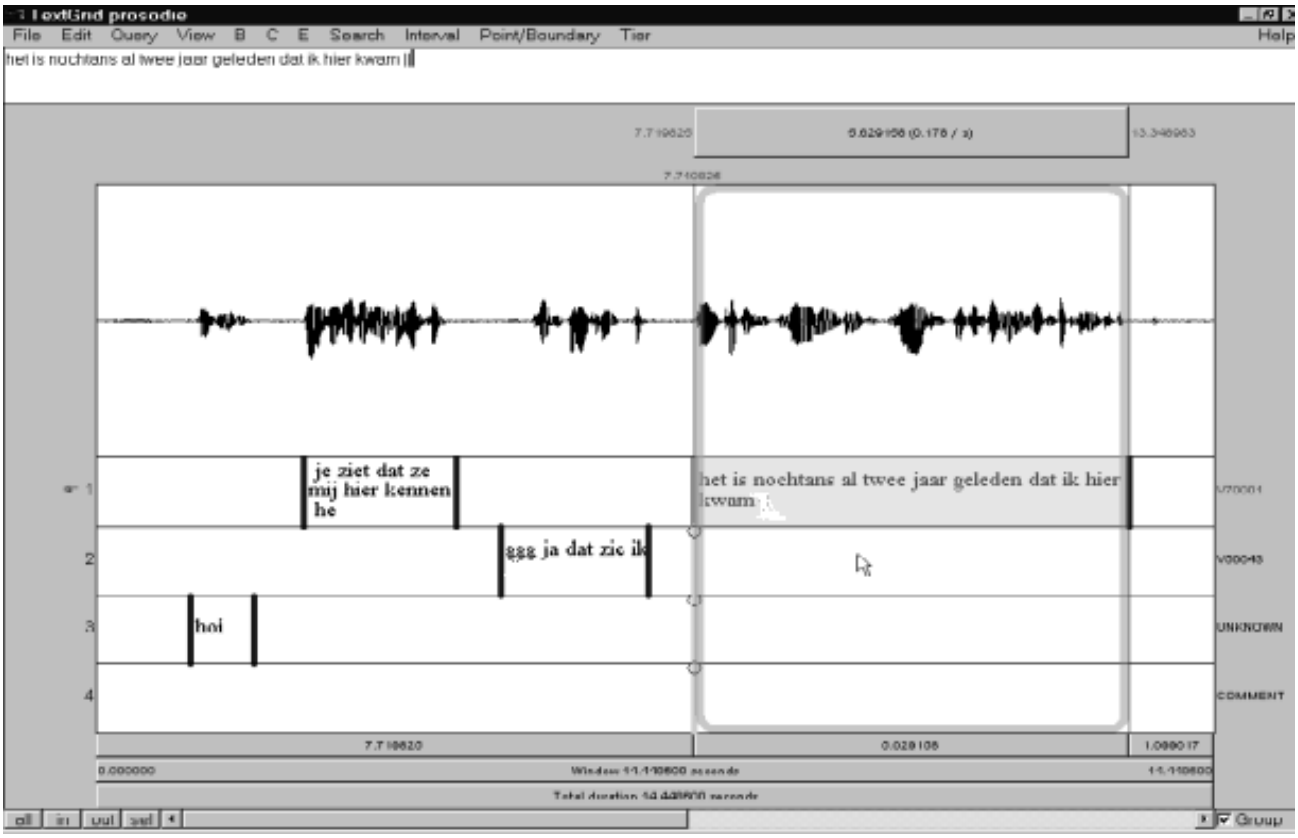

Figure 1. Sample screen of the Praat user interface. A single waveform is shown in the top window. Interactants are represented on separate text tiers (speaker 1, speaker 2, unknown third speaker). Vertical black bars indicate automatically detected strong phrase boundaries (see text).

Ex: *he brought ^ei^ght cases of r^e^d wine*

4. **Segmental lengthening** (symbol '%') is defined as an unusual lengthening of a vowel or consonant that is not accompanied by an auditory impression of prominence or a break. The phenomenon often occurs when the speaker is hesitating, or when s/he is emotionally aroused]. Transcribers were explicitly briefed not to annotate filled pauses as instances of segmental

3.1. Speech corpora

The first part of the pilot study consisted of assembling a test corpus to be annotated. As Dutch (D) and Flemish (F) were expected to be prosodically different (see for instance Gooskens, 1997), two test corpora (TD and TF) of twelve files each were compiled. The speech files were selected from the various main components of the CGN (Oostdijk et al., 2002). The selection was re-

stricted to files with a manually checked word alignment. Only the initial five minutes of each file were included in the test corpus. Three files from each test corpus were selected as a learning corpus; the remaining nine files constituted the test corpus proper (about 8,000 words). The learning corpus comprised read, scripted and unscripted (i.e. spontaneous) speech. In Leiden and Utrecht, i.e. the two Dutch universities, transcribers annotated the speech files of TD, in Antwerp and Ghent, i.e. the two Belgian sites, the TF test corpus was transcribed.

3.2. Learning and annotation

After the transcribers had spent a few days studying the protocol, the experimental part of the pilot study was run in four phases. Everything was done the same way in the two countries/regions (i.e., the Netherlands and Flanders):

Phase 1. The transcribers annotated the first minute of the three learning files, after which the transcriptions were discussed by both transcribers and the site supervisor. On the basis of this feedback, the transcribers corrected their transcriptions, and continued with the following minute of each file. After the second feedback round, they went on to transcribe the entire learning corpus (15 minutes).

Phase 2. As soon as all the transcriptions were available for both sites per country, a so-called mean transcription was derived (see section on evaluation) for each country (i.e. one for TD, one for TF). These transcriptions were checked and corrected by the two site supervisors per country, until a consensus transcription was obtained.

Phase 3. As a last form of feedback, the transcribers were asked to go through the learning materials one more time, and check their transcriptions against the consensus transcription – without making any further changes.

Phase 4. In a period of roughly six weeks, the naive transcribers worked their way through the 45-minute test corpus, without any further supervision.

As soon as all the annotations were available, they were (automatically) checked for formal correctness, and subjected to an evaluation.

3.3. Evaluation phase

The goals of the evaluation were:

(i) to estimate the attainable degree of consistency between students,
(ii) to estimate the time needed to perform the annotations, and
(iii) to make recommendations for the actual production of the annotations.

All the evaluation data refer to the 45-minute test corpora that were processed in phase 4 of the annotation experiment.

The inter-transcriber consistencies for prominence and break strength were quantified by means of Cohen's kappa coefficient (Cohen, 1960). According to Landis & Koch (1977), a kappa between 0.61 and 0.80 points at a substantial consistency. The kappas of all the transcriber pairs are listed in Table 1 (prominence) and Table 2 (break strength). The values above the diagonal are for Flemish and those below the diagonal for Dutch transcribers. The transcribers are indicated by the region (F/D) and the first letter of the site they were working at.

Table 1. Inter-transcriber agreement (kappa coefficients) for prominence annotations. Transcribers from Flanders in upper half of matrix; Dutch transcribers in lower half. U = Utrecht, L = Leiden, G = Ghent, A = Antwerp. Two transcribers per site.

	FG1	FG2	FA1	FA2	
DU1		0.576	0.605	0.603	FG1
DU2	0.633		0.638	0.638	FG2
DL1	0.710	0.589		0.719	FA1
DL2	0.704	0.580	0.592		FA2
	DU1	DU2	DL1	DL2	

Table 2. Inter-transcriber agreement (kappa coefficients) for break strength annotations. Further see table 1.

	FG1	FG2	FA1	FA2	
DU1		0.735	0.695	0.762	FG1
DU2	0.757		0.774	0.768	FG2
DL1	0.738	0.695		0.720	FA1
DL2	0.769	0.732	0.884		FA2
	DU1	DU2	DL1	DL2	

In a similar experiment on prominence labeling by naive listeners, Streefkerk et al. (1997) found kappa values which were typically between 0.45 and 0.60. Our own results are typically better (ranging between 0.58 and 0.72). Although it is hazardous to compare performance across experiments, we would maintain that our transcribers' superior performance is in no small part caused by the use of a standardized protocol and supervised learning stage.

Moreover, table 2 shows that that the kappa values in all the cells for break annotations (considerably) better than the corresponding prominence annotations, to the extent that these kappas are now within the critical range of 'substantial consistency' (see above).

The inter-transcriber differences were also assessed on the basis of simple statistics such as the number of prominences, weak and strong breaks, etc. they indicated. These results are summarized in Table 3 (next page).

Interestingly, all four Dutch transcribers agreed on prominence/non-prominence for 83% of the words. For the Flemish students this was 76%. This difference is not reflected in the kappa values of table 1. This shows that a consistency analysis of pairs of transcribers is not enough and needs to be followed up by a more complete analysis of the consistency across all labelers involved.

A third way of analyzing the data consists in comparing the annotations of each transcriber with a reference that is derived from the mean annotation of the remaining three transcribers that annotated the same data. The mean prominence was the arithmetic mean of the (weighted) transcriber scores: 0 or 1 for prominence, 0 (no), 1 (weak) or 2 (strong) for break strength. The mean prominence

Table 3. Number of words, prominences, strong and weak breaks between and within words, and segment elongations transcribed by each of four Dutch and four Flemish transcribers.

	DU1	DU2	DL1	DL2	FG1	FG2	FA1	FA2
Words	8062	8062	8062	8062	8070	8070	8070	8070
Prominences	1305	889	1439	1519	1768	2084	2289	1964
Between-words								
Strong breaks	698	632	748	744	1009	968	738	1176
Weak breaks	1013	1195	704	835	451	515	201	441
Within words								
Strong breaks	1	2	2	2	2	3	0	0
Weak breaks	5	7	3	4	5	9	5	6
Segment lengthening	14	16	9	7	71	25	26	16

Table 4. Task performance of each individual transcriber, relative to the mean reference performance of the remaining three transcribers within the same country.

	DU1	DU2	DL1	DL2	FG1	FG2	FA1	FA2
Prominences								
Insertions	372	680	127	95	725	568	185	372
Deletions	254	119	349	417	296	392	712	452
Correlation	0.725	0.630	0.789	0.782	0.646	0.671	0.717	0.719
Between-word breaks								
Correlation	0.895	0.876	0.911	0.925	0.909	0.927	0.902	0.912

was 1 if the prominence score was larger than 0.500; the mean break strength was 2 if the mean score was larger than 1.499 and 1 if it was larger than 0.501 (and less than 1.499). Table 4 lists the following data for each transcriber:

(i) the number of prominence deletions/insertions relative to the mean reference,
(ii) the correlation between the individual transcriber prominence scores and the reference score, and
(iii) the correlation between each transcriber's break scores and the reference score.

Some facts that can be derived from table 3 are:

- The Flemish transcribers indicate more prominences than their Dutch counterparts do (25% versus 20% of the words),
- the total number of breaks is very similar across all transcribers, but
- the balance between weak and strong breaks is different between the two regions, and
- within-word breaks and segmental lengthening occur only occasionally.

One transcriber (DU2) seems to have a different view on prominence labeling. One Flemish transcriber (FG1) indicates substantially more segmental elongations and somewhat less prominence than his Flemish colleagues.

From Table 4 it appears that, except for transcriber DU2, Pearson's correlation between individual prominence scores and reference scores emerging from the remaining three transcribers are larger than r = 0.64. In both regions, the number of prominence deletions and insertions relative to the reference is of the order of 40% of the total number of prominences. Correlations between individual transcribers' break scores and the reference break scores are pretty high (typically in excess of 0.9) for all transcribers.

As for the training time required for the transcribers to become proficient in their task, our results indicate that intensive training on three minutes of speech and monitored annotation of another 12 minutes of speech is sufficient to get students without any previous experience or theoretical background in speech prosody to produce prosodic annotations at a level of inter-transcriber consistency that is at least as good, if not better, than that reported in the literature on expert transcribers. The entire training phase (phases 1 to 3 of the pilot) took about 16 hours per student. Once the training phase was completed, students proved able to maintain their level of consistency throughout the pilot experiment. An analysis of the production time needed shows that the eight transcribers worked at a very constant speed (both within and between transcribers) of approximately 40 times real time. That is to say, that it took students 40 minutes of work to provide prosodic annotations for one minute of speech.

4. Discussion and conclusion

In this paper we have presented a relatively simple prosodic annotation scheme for marking a subset of the utterances collected within the Spoken Dutch Corpus with breaks, prominent words and cases of segmental lengthening. These prosodic tags can be provided by non-expert labelers within a reasonable time frame, after they have had a few training sessions with an explicit protocol. The

results of a pilot study reveal that the quality of these labelings, whether measured in terms of the kappa statistic, in terms of percent complete agreement, or in terms of a comparison with a reference labeling, is high, albeit that it is difficult to determine exactly how the results of this fast-and-cheap labeling procedure compare to the results of other approaches reported in the literature, given that these tend to differ from the current one regarding the metrics used to evaluate the transcriptions and regarding the amount of prosodic detail to be transcribed.

It should be pointed out that the pilot study described in the present paper covered a mere 16,000 words out of a total work load of 250,000 words of spoken language to be prosodically annotated. None of the students that were hired during the pilot experiment, are available for the follow-up transcription project, but we are now in the fortunate position that we have at our disposal a sizeable quantity of spoken language together with a consensus 'golden standard' prosodic annotation and a time-tested written protocol of instructions on how to produce prosodic annotations. These assets will enable us, in the near future, to train new generations of student-transcribers of high quality – with only limited input required on the part of our senior staff. Probably, a prosodic annotation monitoring tool can be devised with relative little effort, which can supervise students while going through the training corpus and provide automatic feedback on their performance.

In the pilot study presented above, four labelers for each country annotated identical sets of speech data in order to be able to measure inter-transcriber consistency. Although there is insufficient funding within the CGN to repeat this for the target 250,000 words, multiple labelings of the same material can be exploited by the user of these labels. In fact, preliminary observations suggest that the mean labeling which was automatically derived from four student labelings (see section 3.3), can function as a 'golden standard' for future analyses. Indeed, when this reference was independently checked by the site supervisors, it turned out to closely reflect an 'ideal' labeling of the data. Second, the four parallel labelings can also be used as a basis to compute more gradient prosodic scales to express continuous variation in degree of prominence and boundary strength (e.g. as in Streefkerk et al., 1997). The starting assumption for this would be that more people will agree on stronger breaks and more prominent accents, whereas there will be less consensus on weaker accents and boundaries. It does remain an interesting empirical question, though, whether that newly generated scale does indeed correctly express gradient differences in accent and boundary strength. Be this as it may, the current proposal is that the 250,000-word target sample of spoken Dutch to be prosodically annotated (125,000 Dutch, 125,000 Flemish) be transcribed by two students (for each language variety). These two students would be working at different sites, so as to produce independent labelings for the same materials. This approach would also offer the advantage that differences between the two annotations could be automatically monitored by the site supervisors, in order to pinpoint potential problems with certain transcribers, and to monitor inter-transcriber consistencies.

Obviously, the quality of a labeling scheme does not only depend on inter-transcriber consistency measures and on how much it costs, but also on how useful the resulting annotations are for other research purposes. Even though they are not as rich as the annotations achieved in ToBI (Beckman & Ayers, 1994) or ToDI frameworks, there are reasons to believe that speech data prosodically tagged along the lines sketched above are indeed relevant as a resource for various linguistic studies and for the further development of speech technological applications.

From a linguistic point of view, such data may serve as input for various studies that seek to gain more insight into the different factors that determine why words are accented and why speakers insert prosodic breaks between words, and what determines the variation regarding these two phenomena. For instance, the descriptive statistics from the pilot study presented here suggest that there are interesting differences in the relative frequency of accents and breaks between regional variants of Standard Dutch, provided that these differences are not simply due to different interpretations of the protocol at the different sites. Regarding more technology-oriented uses of the annotated data, it is clear that they are potentially useful both for speech synthesis and speech recognition.

Developers of Dutch text-to-speech systems can use the annotated data as training materials to obtain models that automatically predict accents and breaks in input texts. Next, there is an increasing interest in using prosody for a whole gamut of pre or post-processing tasks in automatic speech recognition and understanding. For instance, there have been recent attempts to use prosodic breaks to re-rank n-best lists of an automatic speech recognizer, to run separate models for words that are accented and those that are not, to first chunk a continuous speech stream into smaller units before it is sent to the recognition module or to automatically punctuate transcribed spoken texts (Chen, 1999). Obviously, in order to make these efforts of integrating prosody into automatic speech recognition and understanding successful, one is in need of vast amounts of speech data that are consistently and reliably marked with prosodic accents and breaks.

5. Acknowledgement

This publication was supported by the project "Spoken Dutch Corpus" (CGN-project), which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government.

6. References

Beckman, M.E. and Ayers, G.M., 1994. Guidelines to ToBI labeling, Version 2.0. Unpublished manuscript, Ohio State University.

Boersma, P. and Heuven, V. van, 2001. Speak and un-Speak with Praat. *Glot International*, 5:341-347.

Boerma, P. and Weenink, D., 1996. Praat. A system for doing phonetics by computer. Report nr. 132, Institute of Phonetic Sciences, University of Amsterdam.

Chen, C., 1999. Speech recognition with automatic punctuation. *Proceedings of Eurospeech (Budapest)*, 447-450.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Physiological Measurements*, 20:37-46.

Gooskens, Ch., 1997. *On the role of prosodic and verbal information in the perception of Dutch and English language varieties*. Doctoral dissertation, Catholic University Nijmegen.

Grover, C., Facrell, J., Vereecken, H., Martens, J.P. and Van Coile, B., 1998. Designing prosodic databases for automatic modeling in 6 languages. *Proceedings of the 3rd ESCA/COCOSDA workshop on Speech Synthesis (Jenolan Caves)*, 93-98.

Gussenhoven, C., Rietveld, T. and Terken, J., 1999. *ToDI, Transcription of Dutch Intonation*. http://lands.let.kun.nl/todi.

Landis, J. and Koch, G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

LDC, 1994. *Switchboard: a user's manual*. http://www.ldc.upenn.edu/readme_files/switchboard.readme.html.

Martens, J.P., Binnenpoorte, D., Demuynck, K., Van Parys, R., Laureys, T., Goedertier, W. and Duchateau, J., 2002. Word Segmentation in the Spoken Dutch Corpus. *Proceedings LREC (Las Palmas)* (this issue).

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J., Moortgat, M. and Baayen, H., 2002. *Proceedings LREC (Las Palmas)* (this issue).

Portele, T. and Heuft, B., 1995. Two kinds of stress perceptions. *Proceedings of the 14th International Congress of Phonetic Sciences (Stockholm)*, 126-129.

Streefkerk, B., Pols, L., and Bosch, L. ten., 1997. Prominence in read aloud sentences, as marked by listeners and classified automatically. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 21:101-116.

Wightman, C. and Ostendorf, M., 1994. Automatic labeling of prosodic patterns. *Proceedings of the IEEE*, 469-481.

Wightman, C. and Rose, R., 1999. Evaluation of an efficient prosody labeling system for spontaneous speech utterances. *Proceedings of IEEE Automatic Speech recognition and Understanding Workshop (ASRU) (Keystone).*