Integrating Linguistic and Performance-Based Constraints for Assigning Phrase Breaks

Michaela Atterer
Institute of Natural Language Processing
University of Stuttgart, Germany
atterer@ims.uni-stuttgart.de

Ewan Klein
Division of Informatics
University of Edinburgh, UK, and
Edify Internat. Development Centre, UK
ewan@cogsci.ed.ac.uk

Abstract

The mapping between syntactic structure and prosodic structure is a widely discussed topic in linguistics. In this work we use insights gained from research on syntax-to-prosody mapping in order to develop a computational model which assigns prosodic structure to unrestricted text. The resulting structure is intended to help a text-to-speech (TTS) system to predict phrase breaks. In addition to linguistic constraints, the model also incorporates a performance-oriented parameter which approximates the effect of speaking rate. The model is rule-based rather than probabilistic, and does not require training. We present the model and implementations for both English and German, and give evaluation results for both implementations. We then examine how far the approach can account for the different break patterns which are associated with slow, normal and fast speech rates.

1 Introduction

Normal spoken language is not delivered in an uninterrupted monotone; prosodic cues such as pauses or boundary tones greatly help the listener to understand an utterance. Most text-to-speech systems use statistical models to find the appropriate locations for prosodic phrase breaks. In this work we use insights gained from the linguistics literature to develop a computational model which assigns prosodic structure to unrestricted text.

We start by briefly reviewing the relationship between syntactic and prosodic structure. Figure 1 shows an example of the right-branching syntactic structure that is standardly assigned to English sentences. Figure 2 shows a much flatter tree which corresponds to widely accepted views of the same sentence's prosodic structure. According to the latter, the Utterance level is partitioned into intonational (I-) phrases,[1] which in turn are partitioned into phonological (ϕ-) phrases. (We ignore lower levels of representation such as prosodic words and syllables for the purposes of this paper.)

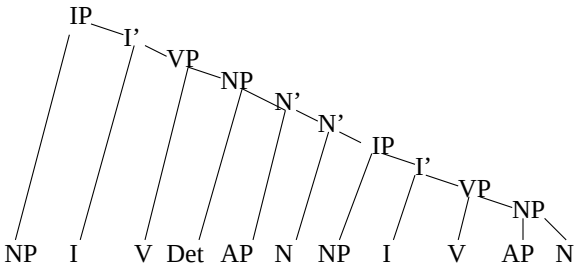

Figure 1: Syntactic structure of a sentence.

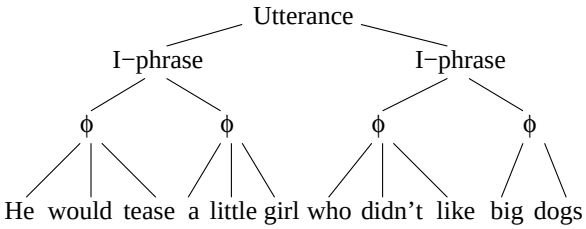

Figure 2: Prosodic structure of a sentence.

In their investigation of the syntax-prosody mapping, Nespor and Vogel (1986) define ϕ-phrases as consisting of a lexical head (e.g., a verb, noun or adjective) together with all the material on its nonrecursive side up until the next head.[2] In the ex-

[1] Intonational phrases are phonologically defined as units which are associated with a characteristic intonational contour; in particular, an I-phrase is marked by the presence of a major pitch accent. The boundary of an I-phrase is canonically manifested as a perceptible pause, accompanied by a local fall or rise in F_0 (fundamental frequency); it can also be marked by constituent-final syllable lengthening, and stronger articulation of constituent-initial consonants.

[2] Here, 'nonrecursive' is intended to cover modifiers and de-

ample of Figures 1 and 2 *tease, little, girl, like, big* and *dogs* are lexical heads. These heads—barring the adjectives—are bundled with the material to their left. The adjectives are included in the same ϕ-phrases as the nouns they modify because they are still inside the maximal projection (NP) of the nouns.

The level of ϕ-phrases can fairly easily be derived from syntax. However, the same is not true of I-phrases. According to the strict layer hypothesis (Selkirk, 1984), an intonational phrase must consist of complete ϕ-phrases. But syntax does not determine how many ϕ-phrases go to make up an I-phrase. To illustrate this point, consider (1), discussed by Gee and Grosjean (1983), where '|' is used to indicate I-phrase boundaries. Both phrasings are acceptable.

(1) By making his plan known | he brought out | the objections of everyone.|

(2) By making his plan known | he brought out the objections of everyone.|

Nevertheless, the ϕ-structure provides a strong constraint on the location of breaks between I-phrases, since an I-phrase can never interrupt a ϕ-phrase.

Although ϕ-structure has been used by others to assign prosodic structure algorithmically (Gee and Grosjean, 1983; Bachenko and Fitzpatrick, 1990), there is no generally accepted method for bundling ϕ-phrases into I-phrases. The main consensus is that I-phrases have "a more or less uniform 'average' length" (Nespor and Vogel, 1986, p.194). In a similar vein, Gee and Grosjean (1983) observe that utterances tend to be split into two or three I-phrases of roughly equal length.

Gee and Grosjean (1983) (and subsequently, Bachenko and Fitzpatrick (1990)) construct I-phrases by comparing the length of the prosodic constituents on both the left-hand side and the right-hand side of the utterance's main verb (or the ϕ-phrase containing the verb), and grouping the verb with the shorter neighbouring constituent. They give little consideration to the grouping of constituents which are not adjacent to the verb. This limitation in their model seems innocuous when dealing with the rather artificially 'well-behaved' set of sentences in their sample. (This 14 sentence corpus, also used by Bachenko and Fitzpatrick, only contains sentences of 11–13 words in length and does not scale up to unrestricted text). However, to be useful in a realistic TTS system our model should robustly run with unrestricted text and not rely – like Bachenko and Fitzpatrick's model – on a correct parser output. Consequently, we need to adopt a different strategy.

terminers as opposed to complements. It is also required that the 'next head' referred to in the definition be outside the maximal projection of the head which forms the basis of the ϕ-phrase.

2 The computational model

Our initial English model was developed within the framework of the LT TTT tokenization toolkit (Grover et al., 2000): this provides a modular and configurable pipeline architecture in which various components incrementally add XML markup to the input text stream. More details of the implementation can be found in (Atterer, 2002). In principle the algorithm consists of two main steps, each of which in turn is broken down into two further steps:

Step 1 Assignment of ϕ-phrases

1. Chunking
2. Restructuring of chunks to build ϕ-phrases

Step 2 Bundling of ϕ-phrases into intonational phrases ("Insert Phrase Breaks")

1. Insertion of breaks using punctuation
2. Insertion of further breaks using balancing and length constraints

The first important step is to identify ϕ-phrases. Although we require some syntactic markup as input to constructing these, a full parse is not necessary. Instead, we carry out a shallow parse using a *chunker*. For English, we use Abney's Cass chunker.[3] Cass builds syntactic structure incrementally starting with a level of simple chunks and then building various levels of more complex phrases above them. Phrases of each level are constructed non-recursively out of constituents of the previous level. For this work we only use the lowest level of units such as `nx` (noun chunk) and `vx` (verb chunk), as illustrated in (3).

(3) <nx>Their presence</nx> <vx>has enriched</vx> <nx>this university</nx> and <nx>this country</nx>, and <nx>many</nx> <vx>will return</vx> <nx>home</nx> <inf>to enhance</inf> <nx>their own nations</nx>.

[3]Cass is available at `http://www.research.att.com/~abney/`

Abney's defininition of chunk is very similar to Nespor and Vogel's notion of ϕ-phrase: "roughly speaking, a *chunk* is the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head, but not including post-head constituents." (Abney, 1996). Chunks defined in this way map almost directly into our ϕ-phrases, except that we also include in the ϕ-phrase any unchunked material on the left boundary of the chunk. For example, the sequence *and* <nx> *this country*</nx> in (3) is converted into a single ϕ-phrase.

For the German version of the model, we used a chunker developed by Helmut Schmid (work in progress) and carried out some subsequent restructuring of the chunker's output. The four main modifications to the chunk structure are as follows.

1 In German, as opposed to English, the auxiliary can be separated from the verb/verb group it belongs to. That is, a complement or modifier can split the verb chunk, and consequently the chunker builds two separate verb chunks. Since the auxiliary does not count as a lexical head, we delete the chunk boundary after it. This is illustrated by examples (4) and (5) where the deletion of the chunk boundary after the auxiliary *hat* results in the ϕ-phrase *hat den Führungsstreit*.

(4) <nx> Der nordrhein-westfälische Ministerpräsident </nx> <nx> Rau </nx> <vx> hat </vx> <nx>den Führungsstreit </nx> <px> bei <nx> den Sozialdemokraten </nx> </px> <vx> kritisiert </vx>. <nx>

(5) <phi> Der nordrhein-westfälische Ministerpräsident Rau </phi>< phi> hat den Führungsstreit </phi><phi> bei den Sozialdemokraten kritisiert. </phi>

2 Proper names, which are often output as separate chunks by the chunker, are attached to a preceding noun. In (5) the name *Rau* has been attached to the preceding noun chunk of (4).

3 Verb particles at the end of sentences are attached to the preceding chunk. Such verb particles are in fact part of verbs, but are sometimes separated from the verb stem, e.g. the particle *auf* from the verb *aufgeben* (*to give up*) in the sentence *Er gab seinen Plan auf.* (Lit: *He gave his plan up.*) In example (7) the particle *ab* is attached to the preceding chunk of (6).

(6) <nx> Die weitere Entwicklung </nx> <px> in <nx> den kommenden Jahren </nx> </px> <vx> hänge </vx> <px> von <nx> den unternehmerischen Qualitäten </nx> </px> <vx> ab </vx>.

(7) <phi> Die weitere Entwicklung </phi><phi> in den kommenden Jahren </phi><phi> hänge </phi><phi> von den unternehmerischen Qualitäten ab . </phi>

4 Phrase-final verb chunks which consist of only one word are also attached to the preceding material. This is also illustrated by (4) and (5) where the final verb chunk consisting only of the past participle *kritisiert* is included in the same ϕ-phrase as the preceding chunk.

After identifying break-options in the form of ϕ-phrases, we have to bundle these constituents into intonational phrases. As mentioned before, there is observational evidence that utterances should be divided into intonational phrases of roughly equal length. Examining the Spoken English Corpus (SEC), Knowles et al. (1996a, p.111) found that speakers insert breaks after about five syllables in most of the cases and that they almost never utter more than 15 syllables without a break.

Our algorithm will thus contain a threshold parameter which sets an upper bound on the length of I-phrases. This value is used to calculate the optimum length of the I-phrases for particular sentences. Even though the threshold sets an upper bound, it is not a rigid one: an I-phrase can become longer in some cases. This is similar to cases in which a speaker would like to pause and maybe take a breath, but has to utter a few more words in order to complete a chunk.

As we mentioned before, we envisage our system as forming one component of a TTS system, and therefore it is reasonable to expect punctuation in the input. This information provides a hard initial constraint on the formation of I-phrases; commas and periods always correspond to I-phrase boundaries. Once we have identified these I-phrase boundaries, the resulting segments are further subdivided by applying the following procedure.

Insert Phrase Breaks

If the number of syllables `ns` in an intonational phrase is greater than threshold `th`, then

(a) Calculate the number of desired breaks `db = ns/th` and the optimum length

`ol` of each new intonational phrase `ol = ns/(db + 1)`.

(b) Determine the location of each new break starting at the beginning of an intonational phrase, counting `ol` syllables forward, and carrying on until the end of the current ϕ-phrase. This is performed `db` times for the obligatory intonational phrase.

So a threshold of 13, for instance, turns the structure shown in example (4) into the one shown in (8) where breaks are marked by '|' and turns the structure in example (5) into the one shown in example (9).

(8) Their presence has enriched this university | and this country,| and many will return home | to enhance their own nations.|

(9) Der nordrhein-westfälische Ministerpräsident Rau | hat den Führungsstreit bei den Sozialdemokraten kritisiert.|

We tried modifying the last step such that the algorithm could return to the beginning of the current ϕ-phrase if this was closer than the end. It is interesting that this obtained slightly worse results, since we believe that the current algorithm is closer to what humans seem to do: reading on until they feel that a break is necessary but not inserting a break until they have completed the current ϕ-phrase.

3 Evaluation Results

We have already alluded to the fact that often there are several equally acceptable possibilities for assigning prosodic structure to a given stretch of text. Consequently, the very notion of evaluating a phrase-break model against a gold standard is problematic as long as the gold standard only represents one out of the space of all acceptable phrasings. Nevertheless, we have adopted the standard evaluation methodology in the absence of a more suitable alternative.

The English model was evaluated using a test corpus of 8,605 words taken from the Spoken English Corpus (SEC) (Knowles et al., 1996b).[4] Our test corpus comprises 6 randomly selected texts from 6 different genres. We calculated recall and precision values. Recall is the percentage of breaks in the corpus that our model finds: recall $= \frac{B-D}{B} \times 100\%$ where B is the total number of breaks in the test corpus and D is the number of deletion errors (breaks which the model does not assign, even though they are in the test corpus). Precision is the percentage of breaks assigned by the model which is correct according to the corpus: precision $= \frac{S-I}{S} \times 100\%$ where S is the total number of breaks which our model assigns to the corpus and I is the number of insertion errors (breaks that the model assigns even though no break occurs in the test corpus). We also calculated the F-score:

$$\mathrm{F} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \times 100\%$$

The results for running the English version of the model with selected thresholds are shown in Table 1. Increasing the threshold decreases the number of breaks that the model assigns: recall goes down, and precision goes up. Decreasing the threshold results in more overgeneration, with recall going up and precision going down. A threshold of 7 produced the best overall results. Reducing or increasing the threshold below 5 or above 12 results in an overall F-score of below 70. However this is not true for certain individual texts. One of the 6 texts we examined was the transcription of a public speech and thus presumably delivered in a different way than news broadcast for instance. (Example 8 was taken from this speech). Its F-score for a threshold of 13 was 71 while its F-score for a threshold of 7 was only 68. Section 4 below contains further discussion of the role played by the threshold parameter in modelling performance.

	Recall	Precision	F-score
th = 4	83	59	69
th = 6	75	66	70
th = 7	73	69	71
th = 8	70	70	70
th = 13	62	79	69
punctuation only	50	92	65
Taylor & Black	79	72	75

Table 1: Results on SEC Corpus

For comparison, the table also shows the results

[4] The SEC is available from `http://www.hd.uib.no/icame/lanspeks.html` and consists of approximately 52k words of contemporary spoken British English drawn from various genres. The material is available in orthographic and prosodic transcription (including two levels of phrase breaks) and in two versions with grammatical tagging.

of two other approaches, namely a baseline model which we ran on our test data and which only assigns breaks at punctuation marks, and Taylor and Black (1998)'s Markov model for English.[5] It should be mentioned that Taylor and Black's model was trained on the SEC corpus, part of which is used for the evaluation here. It is thus optimized for this corpus and has the disadvantage of being less general than our model. Taylor and Black (1998, p.15) report that recall dropped from 79% to 73% when their model was tested on non-SEC data.

	Recall	Precision	F-score
th = 13	93	96	94
Bachenko & Fitzp.	86	89	87

Table 2: Results on Gee & Grosjean Corpus

Table 2 gives the results of running the system against the more homogeneous corpus (14 sentences) of Gee and Grosjean, when restricted to predicting major breaks (intra-sentential and inter-sentential). For comparison, we also show the results reported by Bachenko and Fitzpatrick (1990) from running their rule-based model on the same corpus.[6]

The German version of the model was evaluated using 7,409 words of the news corpus of the Institute of Natural Language Processing (IMS), University of Stuttgart (Rapp, 1998). News broadcasts read by various speakers were hand-labelled with two levels of breaks (Mayer, 1995). For the evaluation we used all breaks without distinguishing between different levels. The results are shown in Table 3. As a comparison, we also show the baseline results using punctuation only, and results achieved by Schweitzer and Haase (2000) using rule-based approaches for German. The first set of results by Schweitzer and Haase were obtained with a robust stochastic parser and a head-lexicalized probabilistic context-free grammar, and the second set by mapping from tag-sequences.

[5]Precision was calculated from the figures in Table 2 on p. 10 in their paper, assuming 1,404 breaks and 7,662 junctures as stated on p. 4 there.

[6]These were calculated from the annotated sentences in their appendix counting major intra-sentential and inter-sentential breaks. Sentences with parsing errors were treated as if no break had been assigned. A relatively high threshold was picked because we only tried to account for major breaks, and thus lower thresholds would cause too many insertion errors.

	Recall	Precision	F-score
th = 4	90	62	73
th = 5	86	66	75
th = 6	84	69	76
th = 7	80	71	75
th = 10	73	75	74
punctuation only	49	93	64
Schweitzer & Haase 1	86	66	75
Schweitzer & Haase 2	71	82	76

Table 3: Results on IMS Corpus

4 Accounting for prosodic breaks at various speech rates

When speakers talk faster they use fewer breaks per utterance, and when they talk more slowly they use more breaks (Trouvain and Grice, 1999). This is reminiscent of what our model does when we increase and decrease the threshold parameter respectively. Intuitively, the algorithm was often able to predict acceptable break patterns for various threshold parameters. The variation in threshold seemed to reflect what speakers would do when varying their speech rate.

In order to capture this effect in a more formal way we tried to evaluate the algorithm on a corpus which was recorded at three different speech rates (Trouvain and Grice, 1999). Three speakers (*CZ*, *PS* and *AT*) read a German text of 108 words 3 times slowly, 3 times at a normal rate, and 3 times at a fast rate.

Trouvain and Grice show that reducing/increasing breaks is not the only prosodic correlate of changing speech rate; for example, speakers also reduce phone durations or pause durations. The extent to which increasing/decreasing the number of breaks correlates with speech rate varies both within and across speakers. One of the speakers, for instance, uses 23 breaks in her first slow version, 28 in her second slow version, and 26 in her third slow version. On average this was definitely more than she used in her normal versions (20, 20 and 24 respectively). To test our algorithm we only used the slow version with the largest number of breaks, the fast version with the smallest number of breaks, and one of the normal

versions which was closest to the average of the normal versions. We did this for each speaker.

We expected to see an effect of the slower version being better modelled by low threshold parameters, and the fast versions by higher parameters. It turned out, however, that the slow versions produced much lower recall/precision values compared to the faster versions. This was due to the fact the when they produced their slow versions, the speakers tended to insert breaks at positions which do not correspond to our ϕ-phrase boundaries, such as immediately after sentence-initial temporal adverbial phrases (which are not marked by commas in German). We would have needed a tagger which distinguishes adverbials of time from other adverbials to account for this. Moreover, further changes in the rules for the restructuring of chunks might have been appropriate, such as preventing breaks before any phrase-final verb chunks up to a certain length. This expedient needs to be approached carefully, however, since when we are trying to model such a small corpus, there is a danger of 'overfitting' the rule set in a way which fails to generalize properly to more extensive corpora.

For the time being, we decided to manually carry out the first step of the algorithm, namely the assignment of ϕ-phrases, in order to test whether the heuristics are useful for modelling different speech rates. We assigned a final ϕ-phrase boundary to all those structural locations where we could find a phrase break in more than one of our 27 spoken versions of the text. This resulted in a structure which could in theory be found automatically if the necessary information was available (e.g. explicitly annotating adverbs of time).

Running the heuristics on this ϕ-structure did indeed show some potential for imitating various speech rates. Table 4 shows recall/precision pairs for running the algorithm with the range of possible threshold values on a slow, normal and fast version by speaker *CZ*. The grey shading in the table shows the best values, i.e. where recall is greater than 90.0% and precision is greater than 80.0.%[7] It does indeed appear that higher thresholds lead to a better model of fast speech rates, and lower thresholds are more appropriate for slow speech rates. The

[7]The model has a general tendency to assign higher recall than precision values. Therefore we have to weigh precision a little bit lower than recall (approximately in a ratio of 8:9) to see the effect. For better readability we leave out the F-scores, which also would only show the effect if weights were included.

threshold	slow	normal	fast
1–3	100.0/82.8	100.0/65.5	100.0/51.7
4	91.7/84.6	100.0/73.1	100.0/57.7
5–7	91.7/88.0	100.0/76.0	100.0/60.0
8	87.5/95.5	100.0/86.4	100.0/68.2
9–10	83.3/90.9	94.7/81.8	100.0/68.2
11–14	79.2/90.5	94.7/85.7	100.0/71.4
15–17	75.0/90.0	94.7/90.0	100.0/75.0
18–21	75.0/100.0	89.5/94.4	100.0/83.3
22–∞	70.8/100.0	84.2/94.1	100.0/88.2

Table 4: Recall/precision values for one slow, one normal, and one fast version of a text read by speaker *CZ*.

threshold	slow	normal	fast
1-3	92.9/89.7	100.0/65.5	100.0/58.6
4	82.1/88.5	94.7/69.2	100.0/65.4
5-7	82.1/92.0	94.7/72.0	100.0/68.0
8	75.0/95.5	94.7/81.8	100.0/77.3
9-10	75.0/95.5	94.7/81.8	100.0/77.3
11-14	71.4/95.2	94.7/85.7	100.0/81.0
15-17	71.4/100.0	94.7/90.0	100.0/85.0
18-21	64.3/100.0	94.7/100.0	100.0/94.4
22-∞	60.7/100.0	89.5/100.0	94.1/94.1

Table 5: Like Table 4 but for speaker *PS*.

threshold	slow	normal	fast
1-3	100.0/72.4	100.0/65.5	100.0/58.6
4	100.0/80.8	100.0/73.1	100.0/65.4
5-7	95.2/80.0	100.0/76.0	100.0/68.0
8	90.5/86.4	94.7/81.8	100.0/77.3
9-10	85.7/81.8	94.7/81.8	100.0/77.3
11-14	85.7/85.7	94.7/85.7	100.0/81.0
15-17	85.7/90.0	94.7/90.0	100.0/85.0
18-21	85.7/100.0	94.7/100.0	100.0/94.4
22-∞	81.0/100.0	89.5/100.0	100.0/100.0

Table 6: Like Table 4 but for speaker *AT*.

tables for the other two speakers (Table 5 and Table 6) show the same tendency. They also reflect the tendency of those two speakers to use the strategy of varying the number of breaks to a lesser extent than *CZ* when speeding up (cf. Trouvain and Grice (1999)).

5 Discussion

Our heuristic can imitate the phrasing of various speech rates. This can be achieved by modifying a threshold parameter. Slow speech rate is imitated by decreasing, and fast rate by increasing this single parameter.

However, the results are not quite satisfactory yet, because some of the steps of the overall procedure for assigning phrase breaks were manually corrected. It would be necessary to implement these additional changes in the chunker rules, and examine whether they enhance or decrease the overall performance. The latter might be the case if they are too genre specific.

As we noted earlier, a more general problem is that larger text corpora for the evaluation of different speech rates are not available. Another approach, which we would like to explore in future work, would be to feed the output of the model into a TTS system and measure human judgements of acceptability.

6 Conclusion

We proposed a model that uses linguistic constraints and a heuristic to assign phrase breaks to unrestricted text. The model does not need any training. This is useful because training corpora marked with intonational phrases are sparse, especially as far as languages other than English are concerned. We show that the model is adaptable to other languages. Its performance is comparable to other phrase break models, and there is still some leeway for improvement. We tested how far a heuristic which is part of the model is capable of capturing changes in speech rate and gained promising results. This is significant given the increasing interest in non-linear modelling of speech rate within the speech synthesis community.

Acknowledgements

We are grateful to Jürgen Trouvain for kindly making his corpus available to us, and to three anonymous reviewers for their comments.

References

Steven Abney. 1996. Chunk stylebook. Available from `http://www.research.att.com/~abney/publications.html`.

Michaela Atterer. 2002. Assigning prosodic structure for speech synthesis: a rule-based approach. In *Proc. of the Speech Prosody 2002 Conference*, Aix-en-Provence.

Joan Bachenko and Eileen Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in english. *Computational Linguistics*, 16(3):155–170.

James P. Gee and François Grosjean. 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT – a flexible tokenization tool. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1147–1154.

Gerry Knowles, Anne Wichmann, and Peter Alderson, editors. 1996a. *Working with Speech: Perspectives on Research into the Lancaster/IBM Spoken English Corpus*. Longman, London.

Gerry Knowles, Briony Williams, and Lita Taylor, editors. 1996b. *A Corpus of Formal British English Speech: The Lancaster/IBM Spoken English Corpus*. Longman, London.

Jörg Mayer. 1995. Transcription of german intonation – the stuttgart system. Technical report, University of Stuttgart.

Marina Nespor and Irene Vogel. 1986. *Prosodic Phonology*. Number 28 in Studies in Generative Grammar. Foris Publications, Dordrecht.

Stefan Rapp. 1998. *Automatisierte Erstellung von Korpora für die Prosodieforschung*. Ph.D. thesis, IMS, University of Stuttgart.

Antje Schweitzer and Martin Haase. 2000. Zwei ansätze zur syntaxgesteuerten prosodiegenerierung. In *Tagungsband der KONVENS 2000 - Sprachkommunikation*, Berlin. VDE-Verlag.

Elisabeth Selkirk. 1984. *Phonology and Syntax. The relation between sound and structure*. MIT Press, Cambridge, Mass.

Paul Taylor and Alan W. Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, 12:99–117.

Jürgen Trouvain and Martine Grice. 1999. The effect of tempo on prosodic structure. In *Proc. 14th Intern. Confer. Phonetic Sciences*, San Francisco.