Syntax and prosody in a Danish Text-to-Speech System

Peter Molbæk Hansen & Ebbe Spang-Hanssen
University of Copenhagen

0. Introduction

This paper is structured as follows:

In section 1. we present the basic structure of a Danish text-to-speech system (henceforth TTS-system), and illustrate some of the special needs of TTS-systems as compared with other natural language processing systems (NLP-systems). In this connection we exemplify some cases of syntactically determined stress loss in Danish.

In section 2. we present and discuss some preliminary results of an empirical investigation of the relation between syntactic structure and pause distribution in read-aloud Danish prose.

In section 3. we describe a specialised grammar formalism for expressing syntactic as well as morphological and phonological structure, and we outline the parsing strategy used to transform input sentences to phonetically consistent and prosodically adequate transcriptions with special emphasis on how to achieve robustness.

1. Syntactic Aspects of the TTS-System

APPENDIX 1 is a schematic overview of the data flow through the serially connected components which converts strings of ASCII-symbols to sound waves in a stepwise fashion. The component labeled SPECIALIZED PARSER (SSPS FORMALISM INTERPRETER) takes care of the lexico-morphological and syntactic analysis needed in order for the system to extract sufficient linguistic information from the input text. This component, henceforth referred to as SSPS (Surface Structure Parsing System), is the one which we hope to be of most interest to computational linguists, and we will not comment on the other components at all, although most of them are of course most important for the text-to-speech conversion process.

The task of parsing text in TTS-systems is different from that of parsing text in systems aiming at assigning semantic structure to input text. APPENDIX 2 highlights some important differences between TTS-systems and other (typically semantically oriented) NLP-systems. The existence of such differences has, of course, rather direct consequences for the design of parsers, since the information to be extracted from the input text is partly different. The task of SSPS is basically to convert a text string representing an input sentence into a representation which contains all phonetically relevant properties of the sentence (prosodic as well as segmental) in string form, i.e. as some sort of systematic phonetic representation. In most older TTS-systems this task is solved

Proceedings of NODALIDA 1991, pages 87-104

by a set of letter-to-phoneme rules, see e.g. Carlson & Granström (1975), which have no (systematic) access to higher-level linguistic information as such. Although several early TTS-systems can produce comprehensible output utterances practically without any syntactico-semantic knowledge, they do not sound "intelligent", and there is much to be gained in synthetic speech quality by introducing such information, since in many cases semantic, pragmatic, and syntactic structure is more or less directly reflected in the prosodic structure of the spoken sentence, determining e.g. the distribution and relative length of pauses, the distribution and relative prominence of stresses, and the intonation contours of whole utterances.

The ideal TTS-system would therefore have to include at least as much linguistic information as the ideal monolingual "text-to-conceptual structure" system. Since, however, a good comprehensibility will suffice for many practical applications of TTS-systems, it seems reasonable to set the level of ambition somewhat lower, and try to see how far syntactic (and of course lexical and morphological) knowledge will get us in the direction of producing intelligent sounding output, before attempting to include semantic and pragmatic information.

In Danish this makes all the more sense, since there are regular patterns of stress loss in nouns and verbs and of potential pause placement at syntactic boundaries which are almost completely predictable from the syntactic context. APPENDICES 3 and 4 illustrate the most important stress removal patterns of this kind.

Note that the conditioning factors as far as stress loss is concerned are signaled in syntactic surface structure: properties such as the definiteness of noun phrases and the transitivity of verbs are easily extractable from text, once morphosyntactic and lexical information is made available through lexico-morphological and syntactic parsing. Note also that the type of stress loss considered here is not a matter of phonetic degree, but a discrete syntactico-phonological matter, viz. absence vs presence of main word stress in the stressable syllable of the words in question. Most languages (including Danish) have regularly unstressed function words such as articles and prepositions, but stress loss in major class words (verbs and nouns) is a pecularity of the Scandinavian languages, the Danish version being probably the clearest example of this phenomenon. This fact in itself clearly necessitates a syntactic analysis.

Needless to say, the actual phonetic manifestations of stressed and unstressed syllables of such words also depend in part on the syntactic context, and this is also true of the rules for placing pauses, but here basic empirical research is needed to determine these rules themselves. This will be the subject of the next section.

2. Pause Placement

2.1 General Principles

It is not surprising that all linguists who have worked on pause location in speech synthesis agree that the good use of pauses highly increases the quality of the speech produced by the machine. This is also our experience. Until now, we have only manually made some pauses in a few sentence representations, but it is quite clear that even a very limited use of pauses makes synthetic speech much more natural.

This means that our work with pause location has a good practical justification, but it should be mentioned that it also is of considerable theoretical interest. The problem concerns the relations between the various components of language:

- pragmatics: discourse structure.
  Do notions like topic and focus play a role in the determination of pause sites?

- semantics: logic structure.
  Does predicate argument structure play a role?
  Are pauses more likely to occur between a
  predicate and its modifier than between a predicate and its arguments?

- syntax: constituent structure.
  Are pauses most likely to be found at important
  phrase boundaries?

- prosody: prosodic units determined by length and stress.
  Is rythm more important than syntax?

At first, it seems reasonable to think that pragmatics and semantics must be the decisive factors. It is a common view that pauses and units of meaning are connected.

This view is reflected, for instance, in the official Danish guide to punctuation. The traditional Danish punctation, whereby we understand here the use of commas, is a grammatical punctuation that Denmark has in common with Germany and the Eastern European countries, whereas the Western European countries, including Norway and Sweden, use a so-called pause punctuation or meaning punctuation.

In the discussion about the commas we clearly see the opposition between on one hand syntax, which is accused of having only poor links to pause location, and on the other hand the meaning, which is connected to the true pause location. However, when the authorities work out the practical rules for pause punctuation, these rules are coined as very syntactic looking rules: put the comma between sentences, in enumerations, around appositions and so forth. At least, syntax is used to indicate where it is possible to put the commas.

As computational linguists we must hope that syntax plays an important role, since syntax is much easier to handle in a machine than semantics, and, as a matter of fact, it turns out that recent research seems to show that purely morphological and syntactic means will be sufficient to make useful pause determination. This does not mean that traditional German-Danish grammatical comma rules are reliable in all respects relevant for pause placement, nor that it is possible to make the ideal pauses without an understanding of the text. What we claim is that it is possible to solve an interesting part of the problems in a purely syntax based system.

2.2 A Short Look at Recent Research

The point of view expressed by J. Bachenko and E. Fitzpatrick (1990) represents one major trend in American research in this field:

Our current analysis rests on two ideas. First it is possible to describe a level of prosodic phrasing that is independent of discourse semantics. Second, this discourse-neutral phrasing depends on a mix of syntactic and nonsyntatic factors; chiefly, syntactic constituency, left-to-right word order, and constituent length. There is no necessary fit between syntactic structure and phrasing, since prosodic phrasing may ignore major syntactic boundaries in order to satisfy the constraints on phrase length. Bachenko and Fitzpatrick, (1990, p. 155).

The authors thus continue the work done by E. Selkirk (1984) and Gee and Grosjean (1983). In their opinion, the most important component of language in this respect is prosody, but prosody, in its turn, is based on syntax, word order and length. This means that syntax is considered as an important, but indirect factor.

At present, we are inclined to think that the American prosodic rules are not fit for our purposes. In order to build the prosodic phrases, the American linguists work in a very buttom-up fashion, starting with very small phonological units. They end up building a prosodic phrase structure which is about as fine-grained as a normal constituent structure. However, if we consider only the observed

pauses, this whole structure seems very hypothetical. By far the majority of the prosodic boundaries they set up are not marked by any pause when people read aloud at a normal speed.

The hypothetical character of these rule systems can be illustrated by the basic rule in Gee and Grosjean (1983), the verb balancing rule, which has been taken over by Bachenko and Fitzpatrick. The verb balancing rule says that the first grouping of major constituents is made around the verb. If the constituent to the left of the verb is short it is grouped with the verb, and a major boundary is made after the verb. On the other hand, if the constituent preceding the verb is longer than the constituent that follows, then a break is made before the verb, and the verb is grouped together with the following unit:

A. (This little incident) // gives (a new zest) (to our investigation)
B. (Chickens) were eating // (the remaining green vegetables)

In this notation, the parentheses indicate the primary prosodic units that have first been identified. The double slash marks the boundary produced by the verb balancing rule.

This rule explains of course why there is often a break before the verb when the preceding phrase is long. But it is impossible to see that there ever is a boundary after the verb when the preceding phrase is short. As a matter of fact, there seldom is a break after the verb. But that is explained in the theory of Gee and Grosjean by the existence of a subsequent rule that builds this first big phrase into still bigger phrases, see APPENDIX 5.

The bigger the constituent, the more important the boundary. The algorithm builds the phrases mechanically in a left-associative way, and the last boundary will always be the most important. Thus, it seems to us that the main effect of the much praised verb balancing rule is to say that there is a break before the verb if the first constituent is long, and that the chances of a new break increase the further away we get from the first break. Bachenko and Fitzpatrick do not hide the fact that it is difficult to observe the verb balancing rule:

During subsequent processing, this balancing effect is usually lost since neither length nor adjacency to a verb play any further role in Gee and Grosjean's analysis (1990, p.162).

Nevertheless, Bachenko and Fitzpatrick stick to the verb balancing rule, to which they propose some amendments, in particular a verb adjacency rule that measures the distance from the verb. They attach so much importance to the prosodic phrasing that they think it rules out clausal constituency. Examples like the following show, they think, that clausal boundaries are unimportant:

Even my fiancee // believes it's only my imagination.

2.3 Our Own Investigations

Our preliminary investigations seem to show that parsing problems play a somewhat more important role than recognized by the American linguists we have mentioned. This difference might be due to the languages we have examined. The Danish rules may be different from the English rules. But our disagreement with these American linguists may also stem from the fact that they identify the syntactic factor with syntactic constituency, whereas we think that the syntactic factor is a matter of how people parse sentences.

Pauses do not serve to indicate the boundaries of syntactic constituents. Perhaps they mostly serve to indicate where complex constituents end, and thus facilitate parsing. Going down into an embedded clause does not cause any parsing problem, but it is often a problem to pop up again at the right place. In parsing terms: it is often more difficult to pop than to push.

We have investigated 3 political reports from the Danish radio, corresponding to 8 pages of written text, and 2 articles from magazines, corresponding to another 8 pages of text. The political reports

from the radio were registered with a tape recorder and later on written down. The two articles were read by a colleague who did not know the purpose of the investigation.

Our main findings are shown in APPENDIX 6. There are boundaries where speakers nearly always make a pause (obligatory sites), and there are others where it is just possible to do it (potential sites). In the latter case, speakers will use the opportunity to make a pause, if the boundary comes a long time after the preceding pause.

We have listed the various cases in an order corresponding to the relative frequency with which the type of boundary in question is used for locating a pause. Pauses occur very frequently around appositions, quotes and the like (37 to 8), somewhat less frequently after a coordinate term that is not an S, etc.

The categories enumerated from 3 to 8 can be subdivided so that some subcategories will get a very high priority. It seems to be the case that there is a pause before a clause if this clause in its turn starts with an embedded clause, as is the case in example 7a below (at når..).

As to PP's, it does not seem to matter that much whether the PP is a complement or a modifier. It is more important how far it is from the head of the construction, i.e. the verb or the head noun in an NP.

The examples listed in APPENDIX 7 illustrate the 8 main syntactic boundaries which we have examined and which are listed above. Very few pauses can not be classified as belonging to any of the 8 categories.

3. The SSPS Formalism and Parsing Strategy

3.1 General Features

Once the rules governing stress and pause placement are established, they must be expressed in a way which is compatible with the "normal" syntactic rules which treat linguistic expressions as strings of terminal symbols in the usual Chomskyan way, and the effects of such rules must be introduced in the parsed output. To this end we use a specialised formalism: SSPS.

An early version of the SSPS formalism is described in Molbæk Hansen (1989) and in Molbæk Hansen (1991). The current version is outlined in Molbæk Hansen et al. (1991). APPENDIX 8 illustrates some of the main features of the current version, which is basically a phrase structure grammar augmented with facilities for expressing restrictions on sister constituents and for expressing feature percolation from daughter nodes to mother nodes. Each rewrite rule has optionally associated with it one or more indented lines expressing such restrictions and/or percolations. The first restriction associated with the rule

NP ::= DET? AP* NOUN

viz. the indented line

NOUN > DEF C "DEFI"

expresses a **restriction**, namely that if the DEF and NOUN constituents are both present, then the values of the feature attribute DEF (definiteness) for DET (DET > DEF) must be compatible (C) with the absolute value DEFI (indefinite)

The last indented line associated with the same rule, viz. the indented line

NP > DEF : "DEFD"

expresses a **percolation**, namely the assignment to the mother node NP of the absolute value

DEFD (definite) of the feature attribute DEF. The reader is referred to Molbæk Hansen et al. (1991) for a fuller description of the notational conventions. In the current version attribute registers whose scope is global within a whole (sub)tree (cf. the hold registers known from e.g. Augmented Transition Networks) are not implemented, but this facility can be easily added, if the need arises.

In SSPS the same formalism is used to express both morphological and surface syntactic constituent structure, and Danish syntax and morphology actually form two sections of the same grammar delimited by a single line of the form

MORF

marking the beginning of the morphological section.

The chart-based parser analyses the input text according to this grammar and its associated morph lexicon, and it schedules its analysis as shown in APPENDIX 9. First each word of an input sentence is analysed in a top-down, first-rule-first fashion by the morphological section of the grammar, and in this morphological mode up to 4 interpretations are accepted for a word. When all words are analysed, the parser mode is switched to syntactic mode, in which the results of the morphological analyses are taken to be terminal edges from the outset. This strategy permits the use of optimization facilities such as precompiled left-corner tables and Kilbury-strategy compatible with bottom-up parsing, cf. Wiren (1987). Note, however, that the parser will only switch mode if the existence of a morphological grammar section is indicated in the grammar. Otherwise, the whole input string (including white space) will be parsed according to the grammar considered as a monolithic whole. Under normal circumstances, however, the grammar prescribed separation of syntax and morphology makes for faster parsing.

3.2 Demarcation of Prosodic Structure

Syntactically conditioned phonological phenomena (like stress loss, pauses, intonation markers etc.) are introduced in the parser output as linearized output symbols of lexical entries matching with zero-length input symbols, which means that they can be introduced everywhere in the string. The technical details of this facility are described in Molbæk Hansen (1991). APPENDIX 10 shows a parse of the input sentence "en sur tjener taber altid ansigt" with stress loss on the finite verb. The situation after morphological parsing is that each word has several interpretations, each with a unique combination of string representation and morphosyntactic feature values (the abbreviations enclosed in < >). In the example the third input word has two noun interpretations viz. the stressed and unstressed variants of the (deverbal) noun "tjener" ('waiter') and two verb interpretations, viz. the stressed and unstressed variants of the present tense verb form "tjener" ('earns'). The unstressed variants have string representations ending in , (comma) and the value STR0 of the feature attribute STR. The syntactic rule

```
S ::= SUBJ VERB ADVPHR* OBJ?
      OBJ > DEF C "DEFI"
      VERB > STR C STR0
      ?
      VERB > STR C STR1
```

states that a sentence may consist of a subject + a finite verb + optional adverbials + an optional object. The restrictions state that if the object is present and has the value DEFI, then the verb must have the value STR0, otherwise the verb must have the value STR1. Such a rule will ensure that the variant tab*0r, will be chosen in the example sentence, and the comma is interpreted by the phonological system as a symbol which prevents stress from being assigned to the preceding word. (For details of the cooperation of the parser component and the phonological transformer component of APPENDIX 1, see Molbæk Hansen (1991)).

3.3 Robustness

A TTS-system must be robust, both in the sense that unsuccesful parses should not be tolerated, and in the sense that the output should be sensible in cases of unidentifiable input. Many modern parsers take care of robustness in a preprogrammed way, typically by chosing the longest partial subcharts in cases where there are unparsed islands in the input, see e.g. Russi (1991). In SSPS **robustness is obtained in grammar prescribed ways.** The grammar writer may base his rules on e.g. his knowledge of frequent constructions. He may then write meaningful rules for such constructions and write **default rules** for configurations not compatible with these rules. Default rules are rules which are matched by any material at the relevant level if they are allowed to apply, that is, if the "structured" rules are not matched. This presupposes a first-rule first strategy.

APPENDIX 11 illustrates the principle, showing a morphological and a syntactic default rule. This principle enables the grammar writer to experiment with various arrangements (orderings) of the same rules and to evaluate the overall performance of each arrangement on a representative sample of test sentences. We are at present engaged in developing such a test sample.

References

J. Bachenko and E. Fitzpatrick 1990: A Computational Grammar of Discourse Neutral Prosodic Phrasing in English, *Computational Linguistics* Volume 16 Number 3 September 1990 p. 155-170.

R. Carlson and B. Granström 1975: A Text-to-Speech System based on a Phonetically Oriented Programming Language, *Speech Transmission Laboratory, KTH Stockholm, Quarterly Progress and Status Report* 1/1975 p. 17-26.

J.P. Gee and F. Grosjean 1983: Performance Structures: A Psycholinguistic and Linguistic Approach. *Cognitive Psychology* 15 1983 p. 411-448.

F. Grosjean and J.P. Gee 1987: Prosodic structure and spoken word recognition, *Cognition* 25 1987 p. 135-155.

F. Grosjean and M. Collins 1979: Breathing, Pausing and Reading, *Phonetica* 36 1979 p.98-114.

F. Grosjean, L. Grosjean and H. Lane 1979: The Patterns of Silence: Performance Structures in Sentence Production, *Cognitive Psycology* 11 1979 p. 58-81.

P. Molbæk Hansen 1989: Syntax, Morphology, and Phonology in Text-to-Speech Systems, *ARIPUC 23 = Annual Report of the Institute of Phonetics, University of Copenhagen* No 23 1989 p. 119-152.

P. Molbæk Hansen 1991: The Linguistic Components of the Danish Text-to-Speech System, *Copenhagen Working Papers in Linguistics* volume 1. 1990/91 p. 153-162.

P. Molbæk Hansen, N. Reinholt Petersen, J. Rischel, and C. Henriksen 1991: Higher Level Linguistic Information in a Text-to-Speech System for Danish, *EUROSPEECH 91 = Proceedings og The 2nd European Conference on Speech Communication and Technology, Genova, Italy, 24-26 September 1991*, Volume 3 p. 1243-1246.

T. Russi 1991: Robust and Efficient Parsing for Applications such as Text-to-Speech Conversion, *EURO-SPEECH 91 = Proceedings og The 2nd European Conference on Speech Communication and Technology, Genova, Italy, 24-26 September 1991* Volume 2 p. 775-778.

E.O. Selkirk 1984: *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge, 1984.

M. Wirén 1987: A Comparison of Rule-Invocation Strategies in Context-Free Chart Parsing, *Proceedings of The Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen, 1-3 April 1987* p. 226-233.

Peter Molbæk Hansen
Institut for Almen og Anvendt Sprogvidenskab
Københavns Universitet
Njalsgade 80
DK-2300 København S

E-mail: pm@cphling.dk

Ebbe Spang-Hanssen
Institut for Almen og Anvendt Sprogvidenskab
Københavns Universitet
Njalsgade 80
DK-2300 København S

E-mail: esh@cphling.dk

APPENDIX 1: THE DANISH TEXT-TO-SPEECH SYSTEM:

ASCII text: De spiste altid middag kl. 19.

TEXT NORMALIZER (A TRIVIAL PREPROCESSING PROGRAM)

Normalized: de spiste altid middag klokken nitten

SPECIALIZED PARSER (SSPS FORMALISM INTERPRETER)

Morphophonemic: di,spis*t0,all+# tidd mIdd-a klokk-0n,nItt-0n

PHONOLOGICAL TRANSFORMER (SPL FORMALISM INTERPRETER)

Phonetic: disbi:sd0'all,tiD?m'edaklCg^nn'ed^n

PHONETIC TRANSFORMER (SPL FORMALISM INTERPRETER)

parameterized:

P1	**P1**	**P1**	**P1**	**P1**	**P1**	**P1**	**P1**	**P1**	**....**	**P1**	**P1**
P2	**P2**	**P2**	**P2**	**P2**	**P2**	**P2**	**P2**	**P2**	**....**	**P2**	**P2**
Pn	**Pn**	**Pn**	**Pn**	**Pn**	**Pn**	**Pn**	**Pn**	**Pn**	**....**	**Pn**	**Pn**

SYNTHESIZER (SOFTWARE SIMULATION)

Wave image: Fr1 Fr2.....Fr 1649 Fr1650

D/A-CONVERTER

Electric oscillations:

LOUDSPEAKER

Sound waves:

APPENDIX 2: TASKS IN TEXT-TO SPEECH SYSTEMS (TTS) AS OPPOSED TO OTHER SYSTEMS (O)

1. TTS: surface structure 1 -> surface structure 2
vs
O: surface structure -> some other structure

Example:
word inventory and order must be identical in input and output in TTS. (TTS-systems involve both analysis (identification) and generation.

2. TTS: Semantic structure is often irrelevant
vs
O: Semantic structure is usually essential

Example: semantic structure essential in translation systems

3. TTS: Prosodic features often essential
vs
O: Prosodic features often redundant

Example 1:
"Jens drak mælk" is distinguished from "Jens drak mælken" both morphologically and prosodically, but for O-parsers the prosodic difference is irrelevant.

Example 2:
In Danish many but not all stød susceptible monosyllabic first parts of compounds exhibit stød-loss. Thus "halgulv" with stød retension and "balsal" with stød loss have a different morphophonological structure. This difference is completely irrelevant in O-systems.

APPENDIX 3: STRESS LOSS IN NOUN PHRASES

(comma => next word unstressed)
(single quote => next word stressed)

Noun phrase internal constructions of various semantic subtypes:

a) Determiner phrase ending in indefinite noun:
ˌet 'stort ˌbundt 'friske 'radiser
to ˌfed 'hvidløg

b) Constructions with indefinite noun + proper name
ˌprofessor Spang-'Hanssen
ˌslagter 'Olsen
ˌtante 'Agate
ˌfætter 'Jens

c) Constructions with indefinite noun + numeral
ˌnummer 'et
ˌindgang '2
ˌtrappe '6

d) Constructions with indefinite noun + "place name"
ˌrestaurant Den 'Gyldne 'Fortun
ˌKap Det 'Gode 'Håb

e) Hypotactic proper name phrases:
ˌJens 'Peter
ˌOle 'Svendsen
ˌJens ˌPeter 'Jensen

APPENDIX 4: STRESS LOSS IN VERB PHRASES

(comma => next word unstressed)
(single quote => next word stressed)

a) Verb with indefinite direct object, irrespective of intervening material:
Jens ‚drikker aldrig 'fadøl
Jens ‚leder efter 'penge

vs

Verb with definite direct object, irrespective of intervening material:
Jens 'drikker aldrig ‚sin 'mælk
Jens 'leder efter 'pengene

b) Verb + adverbial complement, possibly with intervening direct object:
Jens ‚satte sin cykel 'ind i skuret

vs

Verb phrases with peripheral locative adverbial:
Jens 'satte sin cykel 'inde i skuret'

c) A combination of a) and b):
Jens ‚støvede 'bordet 'af

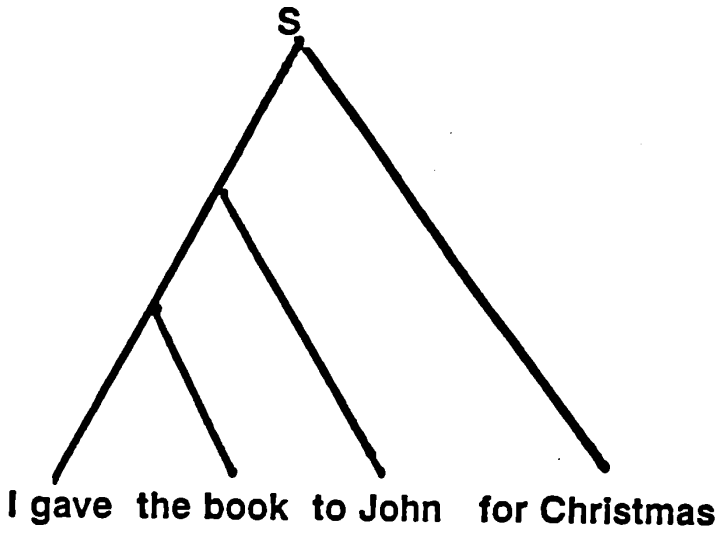
vs

Jens ‚støvede 'borde ‚af

APPENDIX 5: VERB BALANCING RULE

A. (This little incident) // gives (a new zest) (to our investigation)

B. (Chickens) were eating // (the remaining green vegetables)



APPENDIX 6: PAUSES IN WRITTEN DANISH READ ALOUD (16 PAGES)

	YES	NO
OBLIGATORY SITES		
1. After coordinated S	65	5
2. After embedded S	40	2
POTENTIAL SITES		
3. Around appositions, quotes	37	8
4. After coordinated non-S	25	17
5. After complex initial phrase	42	39
6. After complex inverted subject	10	6
7. Before clause	89	75
8. Before PP	41	252
TOTAL NUMBER OF PAUSES	343	
TOTAL NUMBER OF UNUSED SITES		404

APPENDIX 7: SAMPLE PAUSE LOCATIONS

1. Som barn elskede hun kirsebær, // og hun tog imod
posen med en selvfølge, der først bagefter har forundret
mig,// og gik rundt blandt de øvrige passagerer // og bød af sin rigdom. (D,2,1)

2. Men jeg oplevede tre ting den sidste dag, hun levede,// der stadig
fylder mig med undren. (D,1,8)

3. Ved Hitlers magtovertagelse den 30. januar 1933,//marcherede
brunskjorterne,// SA-korpsene //, i et firetimers fakkeltog
gennem porten. (R,3,5)

4. I mange landbrugsegne / er hele landsbyer på flugt// på
grund af mangel på mad og vand // og af frygt for epidemier. (R,6,27)

5. Men ved retten i München // kører en proces mod en af
Wolffs tidligere underordnede,// generalmajor Schutt. (R,5,1)

6. Så skulle tohundredeårsdagen for åbningen af Brandenburger Tor i
dag // være anderledes problemløs.(R,1,13)

7a. En nabo, der ønskede at hjælpe hende,// beklagede
sig over,// at når han kom forbi på vejen / og gerne ville
hilse på hende,// så vendte hun ryggen til. (D,3,11).

7b. Da han erfarede, at jeg endnu ikke havde sovet,//
insisterede han på at give mig en indsprøjtning at sove på. (D,3,28)
Derfor stod Markus Wolff naturligt nok højt på ønskesedlen hos de
vesttyske myndigheder // ved genforeningen. (R,4,13)

APPENDIX 8: SSPS GRAMMAR SECTION

```
NP ::=      DET? AP* NOUN
      NOUN > DEF C "DEFI"
      DET > DEF C AP > DEF
      DET > AN C AP > AN
      DET > AG C AP > AG
      AP > AN C NOUN > AN
      AP > AG C NOUN > AG
      DET > AN C NOUN > AN
      DET > AG C NOUN > AG
      NP > DEF : "DEFD"

DET ::=   QUANT? ART? ENUM?
      QUANT > DEF C ART > DEF
      QUANT > AN C ART > AN
      QUANT > AG C ART > AG
      ART > DEF C ENUM > DEF
      ART > AN C ENUM > AN
      ART > AG C ENUM > AG

QUANT    ::=   WRD
      WRD > W c "WKVA"

ART ::=   WRD? NP
      WRD > W C "WPRON"
      NP > W c "WGEN"
      ART > DEF : "DEFD"
      ART > A : "A"

ART ::=   WRD
      WRD > W C "WDEFR"

ENUM     ::=   WRD
      WRD > W c "WNUM"
```

APPENDIX 9: SSPS PARSER ALGORITHM

```
FOR EACH INPUT SENTENCE
BEGIN
 INPUT A SENTENCE.

 SWITCH TO MORPHOLOGICAL MODE = TOP-DOWN, DEPTH-FIRST,
 FIRST-RULE-FIRST, EACH CHARACTER SURROUNDED BY
 TWO CONSECUTIVE VERTICES.

 FOR EACH WORD
 BEGIN
 PARSE THE WORD ACCORDING TO MORPHOLOGICAL
 PART OF GRAMMAR, ACCEPTING AT LEAST 1 AND
 AT MOST 3 INTERPRETATIONS.
 END
 ARRRANGE THE RESULTS AS TERMINAL EDGES
 IN A CHART FOR THE SYNTACTIC MODE.

 SWITCH TO SYNTACTIC MODE = BOTTOM-UP, LEFT-CORNER,
 EACH WORD-RESULT SURROUNDED BY
 TWO CONSECUTIVE VERTICES.

 PARSE THE SENTENCE ACCORDING TO SYNTACTIC PART OF
 GRAMMAR, ACCEPTING EXACTLY 1 INTERPRETATION.

 OUTPUT INTERPRETATION IN
 LINEARIZED MORPHOPHONEMIC FORM.
END
```

APPENDIX 10: EXAMPLE OF ANALYSIS

Input: En sur tjener taber altid ansigt.

MORPHOLOGICAL ANALYSIS:

0-1 en <WNUM ANUMS AGENC DEFI>
0-1 en, <WDET ANUMS AGENC DEFI>
1-2 surr, <WVIMP STR0>
1-2 surr <WVIMP STR1>
1-2 sur! <WADJ ANUMS AGENC DEFI>
2-3 tjEn=0r, <WSUB ANUMS AGENC DEFI STR0>
2-3 tjEn=0r <WSUB ANUMS AGENC DEFI STR1>
2-3 tjEn+0r, <WVFIN STR0>
2-3 tjEn+0r <WVFIN STR1>
3-4 tab=0r <WSUB ANUMS AGENC DEFI>
3-4 tab*0r, <WVFIN STR0>
3-4 tab*0r <WVFIN STR1>
4-5 all+# tidd <WADV>
4-5 altid, <WSUB ANUMS AGEN DEFI STR0>
4-5 altid <WSUB ANUMS AGEN DEFI STR1>
4-5 altid <WADJ ANUMS AGENC DEFI>
5-6 anh%sIgt, <WVIMP STR0>
5-6 anh%sIgt <WVIMP STR1>
5-6 anh=sIgt, <WSUB ANUMS AGENN DEFI STR0>
5-6 anh=sIgt <WSUB ANUMS AGENN DEFI STR1>

SYNTACTIC ANALYSIS:
en,sur tjEn=0r tab*0r, all+# tidd anh=sIgt

OUTPUT FROM PHONOLOGICAL TRANSFORMATION:
ens'u:?rtj'E:nCtæ:bC'al?,tiD?'an,segd

APPENDIX 11: ROBUSTNESS IN THE SSPS-SYSTEM

UNSUCCESFUL PARSES ARE AVOIDED BY
1) FIRST RULE FIRST STRATEGY
2) DEFAULT RULES:

MORPHOLOGICAL DEFAULT RULE:

STEM ::= letr+
STEM > M : "MNSAC MCC0 MNPX"

I.E. ANY LEXICALLY UNIDENTIFIABLE
PART OF AN INPUT WORD WILL
DEFAULT TO
1) A NOUN STEM,
2) OF THE COMMON GENDER,
3) WITH PLURAL IN -ER,
4) WITHOUT -E- OR -S- AS 1ST PART OF COMPOUND
5) WITH SPELLING = MORPHOPHONEMIC REPRESENTATION.

E.G. AN INPUT WORD LIKE "SMULPERNE"
WILL BE OUTPUT AS:

"smulp+0rn0" <WSUB ANUMP DEFD>

SYNTACTIC DEFAULT RULE:

S ::= WRD+

I.E. "A SENTENCE IS A SEQUENCE OF WORDS"

THIS PRINCIPLE ALLOWS ANY DEGREE OF MORPHOLOGICAL
AND SYNTACTIC EXPLICITNESS TO BE COMBINED WITH
A GIVEN PRACTICAL IMPLEMENTATION.