GENERATING PROSODIC STRUCTURE FOR RESTRICTED AND "UNRESTRICTED" TEXTS

*Anders Lindström** *Merle Horne*** *Tomas Svensson****
*Mats Ljungqvist** *Marcus Filipsson***

** Telia Promotor Infovox AB, P.O. Box 2069, S-171 02 Solna*
E-mail: {anders.lindstrom|mats.ljungqvist}@infovox.se
*** Dept. of Linguistics, Lund University, Helgonabacken 12, S-223 62 Lund*
**** Dept. of Linguistics, Stockholm University, S-106 91 Stockholm*

ABSTRACT

A new text analysis component for the generation of prosodic structure in Swedish text-to-speech conversion is described. It combines shallow syntactic analysis and prosodic parsing algorithms with referent tracking and treatment of lexicalized phrases. It is explained how this approach can be used in restricted and unrestricted texts.

INTRODUCTION

Text-to-speech (TTS) conversion is performed in three stages. In the first stage, the text is analyzed and a symbolic linguistic representation, including pronunciation and prosodic structure, is produced. In the second stage, this symbolic representation is realized in terms of a synthesis specification[1], and in the third and last stage, this specification is used to drive the synthesizer.

One of the most important functions of the first stage of TTS conversion is to produce a proper prosodic description. Such a description of an unrestricted text passage cannot be derived from syntactic information alone, but also depends on semantic and pragmatic factors. A current trend is to use shallow syntactic analysis, such as that produced by applying prosodic grouping algorithms to the output of a part-of-speech tagger. For "discourse-neutral" English texts, it has been shown to be possible to obtain improved prosodic phrasing using this approach [1].

[1]The specification can typically be the formants, bandwidths, amplitudes etc. of a formant synthesizer, or unit selection information plus prosodic parameters of a concatenative synthesizer.

In previous work [2], we have shown that an important aspect of text analysis for TTS conversion, at least in restricted texts, involves keeping track of the coreferential status of lexical items, so that already mentioned items or concepts can be deaccented when re-encountered in the text.

In this paper, we propose a new text analysis component (TAC) for the generation of prosodic structure in Swedish TTS conversion. The structure is a hierarchy of the prosodic constituents "prosodic word", "prosodic phrase" and "prosodic utterance", coupled with information regarding coreferential status (affecting degree of prominence) and boundary strength (later realized as boundary tones, degree of final lengthening and pause length) [3, 4]. We combine shallow syntactic analysis, referent tracking, prosodic parsing algorithms and treatment of lexicalized phrases, and show how this approach can be used both in restricted and unrestricted texts.

The structure of the TAC will be outlined, and the tagging and referent tracking modules will be examined. This work is based on what has previously been reported [2, 3, 5], and the overall goal is to produce a system for high-quality text-to-speech conversion not only in Swedish, but also in other languages [6].

TEXT ANALYSIS

As pointed out earlier [5], the first stage of TTS conversion should not be regarded as a simple "text pre-processor", but rather as a quite complex knowledge-based system, with many highly specialized knowledge sources, that are all located *within* the TTS system, and are allowed to share knowledge with each other. Several linguistic levels of analysis are required to produce the desired prosodic structure. Pragmatics, semantics, and syntax are necessary components in such a scheme.

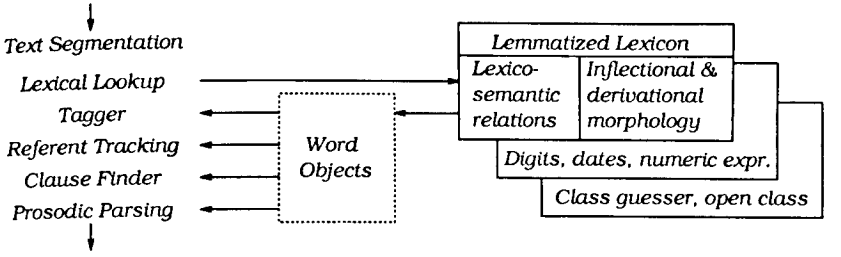

Figure 1: *System architecture of the text analysis component* (TAC)

An outline of our system structure is shown in Figure 1. First, the input text is segmented into sentences and words. The words, or word sequences, as is the case when dealing with lexicalized phrases, are looked up in the lexicon, which is an implementation of a large, lemmatized pronunciation dictionary [7], further equipped with lexico-semantic relations. The result of lexical lookup is stored in word objects in terms of morphological and semantic tags. The sequence of word objects is disambiguated by the tagger. This is a hybrid tagger, where rule-base constraints, dealing with lexicalized phrases, are applied both before and after probabilistic tagging. The resulting, disambiguated sequence of word objects is accessed by a module for referent tracking [2], which makes use of a memory function combined with the lexico-semantic relations stored in the dictionary, to keep track of and mark already mentioned items or concepts. The resulting structure of words, tags and "new/given" status information is used by a pattern matching component to delimit clauses. This information is used, then, to construct prosodic phrases, within which prosodic words are defined [3].

Part-of-speech tagging

The probabilistic part of the tagger used for part-of-speech (POS) disambiguation is an adaptation to Swedish of the Xerox POS tagger [8], a first order Hidden Markov Model based tagger. A tagger of this type has a limited scope of two tokens, making decisions based on transition and symbol probabilities. The texts used for training have been derived from the Stockholm-Umeå Corpus (SUC) [9]. The tagset used is a subset of the SUC tagset consisting of 43 tags.

Choice of tagset content and size is crucial to performance, since the tagset provides the description of the language, with which the tagger is to detect distributional distinctions. We have found that tagset content should reflect as many of the distributional distinctions as possible, and yet not reflect ambiguities that the model cannot resolve, given its inherent limitations.

When constructing a tagset to allow the shallow syntactic analysis required, we found that the language description obtained using a small tagset is not necessarily the best for resolving ambiguities resulting from applying this tagset to a given language. Improvements can be made

Table 1: *Tagging accuracy when A: Tagged & evaluated with 27 tags, B: Tagged & evaluated with 43 tags and C: Tagged with 43 tags & and evaluated when mapped onto the 27 tag subset.*

	Correctly tagged tokens [%]		Tags per word
	Total	Ambiguous	
A	95.1	86.7	1.65
B	94.8	86.7	1.68
C	96.1	90.2	1.65

by splitting and relabeling original tags, in order to separate lexical items with conflicting distributional properties, while avoiding excessive increase in perplexity. The results shown in Table 1 demonstrate the considerable improvements in tagging accuracy that can be achieved by tagging with a larger set (43 tags) and mapping onto a smaller set (27 tags).

Referent tracking

We have previously shown that, when applied to texts from a known, restricted domain, referent tracking algorithms can benefit from taking into account several types of lexico-semantic relations that may indicate co-reference, e.g. morphological identity (both inflectional and derivational), (partial) synonymy, hyponomy-hyperonymy ("is-a" relations) and meronymy-holonymy ("has-a" relations) [2]. In addition to this, some concepts can be considered as pragmatically "given" in a restricted domain, such as *per cent* in stock market texts. When confronted with a word in running text, the referent tracking algorithm consults its "memory"[2] of processed entities to determine whether the current word has been recently mentioned or not, and assigns a status of either "given" or "new" to the word.

Restricted vs. "unrestricted" text

By "unrestricted" texts we mean texts, which can in fact consist of sections, which are narrow in domain, but where these domains and the section boundaries are *not known* to the TTS system in advance. Because of the modular architecture used [5], going from a restricted domain to either another restricted domain, or to unrestricted text, is achieved by inserting or changing knowledge only in a few well-defined places in the system.

For better coverage on unrestricted text, more effort has to be spent on modelling different textual conventions, such as date formats, fractions, abbreviations etc. This is achieved by extending the regular grammar that performs text segmentation and tokenization on the one hand, and, on the other hand, adding to the lexicon corresponding "methods" dealing with those conventions. In this way, the form and internal structure of dates, e-mail addresses etc. can be exploited, and dealt with using (sometimes) more appropriate methods than lexical listing.

As regards lexico-semantic relations, necessary for referent tracking, in the unrestricted case it is possible, even without knowledge of topic structure (see Discussion), to recognize coreference using morphological identity relations from the lemmatized lexicon.

The identification of lexicalized phrases in unrestricted texts [10], which is important both in order to increase the naturalness of synthetic speech and to ease the burden on the probabilistic tagger [5], has been taken even further in a restricted domain than is otherwise possible: When applying the TAC to texts from the stock market domain [3], lexical items are marked with semantic tags such as share-name (*Atlas Copco, Ericsson* etc.), share-type (*A* or *B*) and share-modifier (*bundna (bound)* or *fria (free)*). These tags are later used in the first stage of tagging, to match patterns, e.g. share-name share-type share-modifier, and construe lexicalized phrases, like *Atlas Copco B fria*, which behave like lexical items both grammatically and prosodically.

When it comes to POS tagging, a Hidden Markov Model based tagger, such as the Xerox tagger, is especially easy to adapt to new domains, since it only requires raw, untagged text, a tagset and a lexicon to be retrained.

DISCUSSION

In our work on a restricted domain, the lexico-semantic relations were coded manually on top of the lexicon, but results from the fields of lexicography and information retrieval indicate that lexico-semantic relations could be automatically inferred from other sources, such as corpora, "ordinary" dictionaries, or thesauri. While thesaural information can be directly useful for restricted texts, its use may be a little less straightforward in unrestricted texts, because of problems with polysemy. A solution to this could be to use word sense disambiguation techniques [11]. Related techniques [12] could be used to find the topic (or section) boundaries that are not known in advance, but which are relevant in order to determine a better domain for the referent tracking algorithms.

Different parts of this system place different demands on tagger output, sometimes conflicting with the tagset required by the tagger for optimum accuracy. In our experience, it is virtually impossible to meet the tagger-external demands, while also meeting tagger-internal ones. Therefore, for the sake of tagging accuracy, one should choose a tagset that meets the internal demands only, and leave unresolvable ambiguities to other modules.

Further work is needed in several of the areas mentioned in this paper. Particularly, the area of lexicalized phrases needs to be further studied from both grammatical and prosodic points of view. The proposed text analysis component (TAC) also needs to be formally evaluated.

ACKNOWLEDGEMENT

We thank Telia Research and the HSFR/NUTEK Language Technology Programme for supporting this work, Gunnar Eriksson, Dept. of Linguistics, Stockholm University, for his work on modifying the tagger for Swedish, and Janne Lindberg, Dept. of Linguistics, Stockholm University, for sharing his results on lexicalized phrases.

REFERENCES

[1] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16:155–167, September 1990.

[2] M. Horne, M. Filipsson, M. Ljungqvist, and A. Lindström. Referent tracking in restricted texts using a lemmatized lexicon: Implications for generation of intonation. In *Proc. of the European Conf. on Speech Technology*, volume 3, pages 2011–2014, Berlin, 1993. ESCA.

[3] M. Horne and M. Filipsson. Generating prosodic structure for Swedish text-to-speech. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 711–714, Yokohama, 1994.

[4] M. Horne and M. Filipsson. Computational modelling and generation of prosodic structure in Swedish. In *Proc. of the 13th Intl. Congr. of Phonetic Sciences*, Stockholm, 1995.

[5] A. Lindström and M. Ljungqvist. Orthographic processing *within* a speech synthesis system. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 1683–1686, Yokohama, 1994.

[6] M. Ljungqvist, A. Lindström, and K. Gustafson. A new system for text-to-speech conversion, and its application to Swedish. In *Proc. of the 3rd Intl. Conf. on Spoken Language Processing*, pages 1779–1782, Yokohama, 1994.

[7] P. Hedelin and D. Huber. A new dictionary of Swedish pronunciation. In Kjell Morland and Kari Sørstrømmen, editors, *Proc. of the Scandinavian Conf. in Computational Linguistics*, pages 105–117. Norwegian Computing Centre for the Humanities, Bergen, Norway, 1991.

[8] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proc. of the 3rd Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy, 1992.

[9] E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. The linguistic annotation system of the Stockholm–Umeå corpus project. Technical report, Dept. of General Linguistics, Umeå 1992.

[10] J. Lindberg. Detektering av lexikaliserade fraser för text-till-talkonvertering. In *Proc. of The 9th Conf. of Nordic and General Linguistics*, Oslo, 1995. Forthcoming.

[11] D. Yarowsky. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. of the 14th COLING*, pages 454–460, Nantes, France, 1992.

[12] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21, 1991.

[2] The memory size is rather arbitrarily chosen to be the most recent 60 words in the input text.