Integrating Syntagmatic Information in a Dictionary for Computer Speech Applications

Dieter Huber
Chalmers University of Technology

Abstract

Conventional dictionaries, albeit they often comprise an impressive amount of *paradigmatic* information on various aspects of linguistic description, usually pay only little attention to the representation of *syntagmatic* information. Admittedly, apart from spelling conventions and rules of inflectional agreement, the co-occurrence of individual lexical items will not normally change the orthographic shape of a word when it appears in written text. In spoken language, however, the phonetic realization of words is heavily influenced by context and may change dramatically in a variety of ways, including segmental as well as prosodic features. These changes need to be taken into account in both computer speech synthesis and automatic speech recognition. In this paper, therefore, we argue for the inclusion of syntagmatic information in dictionaries which are developed for the special purpose of spoken language processing in computer speech applications. Two kinds of syntagmatic information will be considered in more detail: *Case Frames* and *Collocations*.

1. Introduction

Spoken language differs from written language in several important respects. For one, natural human speech does not normally present itself in the acoustical medium as a simple linear string of discrete, well demarcated and easily identifiable symbols (i.e. representing the letters and signs of some specified alphabet) and blocks of symbols (i.e. representing individual words separated by blanks), but constitutes a continuously varying signal which incorporates virtually unlimited allophonic variations, assimilations, reductions, elisions, repairs, overlapping segmental representations, grammatical deficiencies, and potential ambiguities at all levels of linguistic description. There are no "blanks" and "punctuation marks" to define words or indicate sentential boundaries in the acoustical domain. Important components of the total message are typically encoded and transmitted by non-verbal and even nonvocal means of communication. Syntactic structures, at least in spontaneous speech, are often fragmentary or highly irregular, and cannot be described in terms of established grammatical theory.

Given these differences, clearly, the computational models, tools and techniques developed for natural language processing (NLP) of written material are not immediately and automatically applicable to spoken language processing (SLP) of human speech. In particular, SLP for practical computer speech applications such as for instance text-to-speech synthesis (TTS) and automatic

Proceedings of NODALIDA 1991, pages 133-139

speech recognition (ASR) requires lexical information that is not normally contained in conventional (including machine-readable and machine-tractable) dictionaries.

In an earlier paper (Huber 1989) a speech parsing algorithm has been presented which exploits the prosodically cued chunking present in the acoustical speech signal and uses it to perform speaker-independent segmentation and broad classification of continuous speech into functionally defined information units. A fundamental objective associated with this approach is to integrate speech signal processing and natural language processing techniques (both linguistic and stochastic) in order to fully exploit the combination of partial information obtained at various stages of the analysis. The contents and structure of the lexicon component to be used with this SLP parsing system have been described in (Huber 1990). In that paper, the overall format of the lexicon has been defined as a medium-sized Swedish monolingual pronunciation dictionary incorporating four kinds of lexical information:

- *phonological information*, i.e. narrow phonetic transcriptions reflecting both standard pronunciation usage and principle variants;
- *paradigmatic information* on various aspects of linguistic analysis specially relevant for SLP purposes;
- *syntagmatic information* reflecting the use of words in context;
- *statistical information*, i.e. data on the frequency of occurrence in a large corpus of language material.

The representation of phonological and paradigmatic information is described in more detail in a separate paper published in this volume (Hedelin & Huber 1992). In this paper I shall focus on the use of syntagmatic information in spoken language processing for practical computer speech applications such as speech synthesis and automatic speech recognition.

2. Syntagmatic Information

Words are normally applied in context, i.e. they co-occur with other words preceding and following them, with which they combine into larger structures like phrases, clauses, and sentences to express the intricate meanings of language. These co-occurrence relations are linguistically predictable to a greater or lesser extent, and can be specified in either grammatical or lexical terms as tactic (context-sensitive) rules and restrictions. For instance, adverbs co-occur with verbs (but not necessarily the other way round), transitive verbs require a direct object as complement (at least), and *kill* collocates with *animal, poison, victim*, etc, but not with, say, *sky, paper* and *blue*.

Conventional word dictionaries, albeit they often comprise an impressive amount of *paradigmatic* information on various aspects of linguistic analysis, pay only little attention to these kinds of *syntagmatic* relationships that exist between individual lexical items. In as far as co-occurrence data are included in dictionaries at all, they are usually found among the "examples of usage" listed under one or another of the component lexemes, i.e. stated implicitly, randomly, and without any claims of consistency and comprehensiveness. This imbalance between paradigmatic and syntagmatic information in most of today's dictionaries reflects not only the "division of labour" commonly assumed to exist between lexicographers on one side and grammarians on the other. Even more so, it reflects the traditional preoccupation of both lexicographers and grammarians with written language. Admittedly, apart from spelling conventions (e.g. sentence initial capitalization, use of a hyphen in certain compounds) and rules of inflectional agreement (e.g. he *works* versus they *work*) valid in some of the worlds languages, co-occurrence relationships will not normally change the orthographic shape of a word as it appears in written text. In spoken language, however, the phonetic realization of individual words is heavily influenced by context (both co-textual and situational) and may change dramatically in a variety of ways, including segmental (e.g. assimilation across word boundaries, reduction, elision, sandhi) as well as prosodic (e.g. speech rate, duration, pausing, intonation, accentuation, stress) features. For example, the actual pronunciation of the Swedish word:

NATURLIGTVIS
(naturally)

which in isolated (lexical) usage may be transcribed phonemically as:

/nɑ'tɯrligt'vis/

and in a narrow phonetic transcription as:

[na'tɯ:ɭɪt'vi:s]

may be realized phonetically in informal conversation as:

[n'aɩɵs]

or as:

[n'aɵs]

or even as:

[hanɦɑ:ɳaɩɵs`m̩ɔŋ:a]

when it appears in the context:

HAN HAR NATURLIGTVIS MÅNGA...
(He has naturally many...)

thus involving not only various kinds of assimilation across word boundaries, but also a shift of accentuation with concomitant reductions and elisions of the unstressed syllables, and an increase of duration in both the stressed and in the phrase-final syllables.

Clearly, this kind of information about the phonetic variability of words in different kinds of co-textual and situational context is of paramount importance for all aspects or speech signal processing (analysis, synthesis, transmission, coding, compression, enhancement, etc) and computer speech technology (text-to-speech, speech recognition, speaker identification and verification, etc). For instance, text-to-speech systems using standard syntactic parsers designed to find "major syntactic boundaries" at which cross-word coarticaluation needs to be interrupted and the intonation contour has to be broken into separate units that help the listener to decode the message, invariably come up with the same two kinds of problems: (1) they tend to produce not one (the most probable, semantically most plausible) but several alternative parses, and (2) they produce too many boundaries at falsely detected or inappropriate sentence locations (e.g. Klatt 1987). Perceptual evaluation of these synthesized contours reveals that listeners get distracted and often even plainly confused by too many, prosodically and/or segmentally marked boundaries, while too few breaks just sound as if the speaker is simply talking too fast. This shows not only that the amount of segmentation and the correspondence between syntactic and prosodic units are dependent on the rate of speech, but also that listeners apparently neither expect, nor need, nor even want prosodically cued information about all the potential richness in syntactic structure described by modem syntactic theories, in order to decode the intended meaning of an utterance.

In order to be able to handle these kinds of phenomena in practical computer speech applications, we propose to include co-occurrence information into our Swedish pronunciation dictionary, which has been designed to provide continuous lexical support to the language processing components in text-to-speech synthesis and automatic speech recognition. In the following two sections, two kinds of co-occurrence data will be discussed in more detail: *caseframes* and *collocations*.

3. Case Frames

The grammar formalism adopted for the SLP parsing algorithm presented in Huber (1989) is based on Fillmore's case grammar (Fillmore 1968). The following reduced set of cases for verb entries has been adopted from Stockwell, Schachter and Partee (1973):

AGENT	- animate instigator of the action
DATIVE	- animate recipient of the action
INSTRUMENTAL	- inanimate object used to perform the action
LOCATIVE	- location or orientation of the action
NEUTRAL	- the thing being acted upon

According to this approach, a caseframe is thus defined as an ordered array composed of the entire set of cases:

```
caseframe = array[agent...neutral]
```

in which each case can be either required (req) or optional (opt) or disallowed (dis), and must be marked accordingly.

Since different verbs often share the same particular kind of caseframe, we propose to store the entire set of 3^5 logically possible caseframes as an indexed list, using the indices as pointers (identifiers) with the respective verb entries in the lexicon. Thus, instead of listing the complete case-frame specification together with the lexical entry, as in the following example for the Swedish verb "hacka" (to chop):

```
hacka 3    type: verb
           infl: v1
           freq: 4
           tran: [2hak:a]
           case: agent - req
                 dative - dis
                 instrumental - opt
                 locative - opt neutral - opt
```

we propose to use the indexed representation format, which results in the more space-economic and search-effective structure:

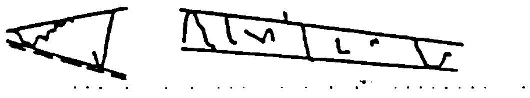
```
hacka 3    type: verb
           infl: v1
           freq: 4
           tran: [2hak:a]
           case: 97
```

Further work in sematic case frame representation is presently directed towards the extension of individual case states marked as with "req" or "opt" lexical hypotheses derived from KWIC-studies of coherent speech.

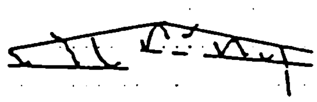
4. Collocations

Collocations constitute the recurrent combinations of words as they appear in context. These combinations include both phraseologically relevant collocations such as *när det gäller att* (Eng: as far as ... is concerned) and phraseologically irrelevant collocations, like *och han* (Eng: and he). While only the former, i.e. phraseologically relevant class of collocations has so far attracted some interest in the linguistic research community, both groups are interesting and important from an information theoretical point of view.

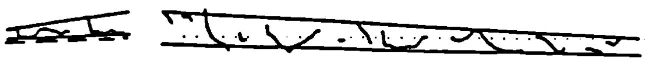
In a sense, collocations can be seen as occupying the border area between grammatical and lexical descriptions of language. In most of today's NLP applications they are handled by the syntactic parsing component with little or no support from the lexicon. We argue, on the other hand, that collocations are primarily a lexical phenomenon that can be handled more effectively by the lexicon component and thus should be included in the dictionary. After all, it makes little sense and would indeed be extremely uneconomical for real-time computer speech applications, to generate these recurrent constructions over and over again from their component words. Also, there is convincing evidence from psycholinguistic research indicating that collocations are indeed part of the mental lexicon that a speaker has at his or her disposal. In natural human speech, collocations are normally processed as coherent lexical blocks equivalent to prosodic phrases (cf. Selkirk 1984), i.e. signaled by a high degree of coarticulation between their component words and an integral melodic pattern. This latter tendency has been found to be particularly prevalent in sentence initial collocations, sentence final collocations, and idioms (i.e. collocations whose collective meaning can not be derived on the basis of the meanings of its component words). Figure 1 below shows examples from our Swedish speech material (cf. Hedelin & Huber 1990) which illustrate the effects of prosodic marking for sentence initial collocations.



Text 1; Sentence 30; Speaker BMH (female)
Dagen därpå undertecknades i klostret ett fördrag...



Text 1; Sentence 30; Speaker LR (female)
Dagen därpå undertecknades i klostret ett fördrag...



Text 2; Sentence 24; Speaker LO (male)
Men å andra sidan har 40 procent av städernas och köpingarnas...

Figure 1. Intonation contours for the sentence initial collocations

Based on these findings, we propose to incorporate the most frequently occurring collocations in the dictionary, including both their orthographic shape and narrow phonetic transcription(s). Special attention is directed towards a complete and comprehensive representation of sentence initial, e.g.

TROTS DETTA

[trɔts`d̡ɛ̡t:a]

(despite of that)

sentence final, e.g.

ÅR SEDAN

['o:s̡ɛ̡n]

(years ago)

and idiomatic constructions, e.g.

I OCH FÖR SIG

[i:ɔfœ's̡ɛ̡j]

or

[i:fœ's̡ɛ̡j]

(as a matter of fact)

both because of their recurrent status as prosodically cued blocks, and their high potential for the kind of expectancy-driven island parsing approach advocated in (Huber 1989). Consequently, collocations are listed under each component head word, which enables the parsing system to effectively guide the word segmentation and identification process by generating expectations resulting from partial linguistic analyses.

Complete listings of the Swedish collocations in descending order of frequency, based on statistical analysis of a one-million-words corpus of news texts (cf. Allén 1975), have been obtained from the Department of Computational Linguistics, University of Gothenburg. These data are stored in machine-readable format which allows direct transfer and incorporation of the orthographic representations into the pronunciation dictionary.

In addition to these listings, we also propose to include two statistical measures: (1) the observed frequency (F_{col}) for each collocation, i.e. indicating the number of occurrences in the accumulated corpus material, and (2) the *constructional tendency* (F_b), i.e. a measure which indicates for each component word its tendency to be bound in collocational constructions. F_b thus represents the ratio between the instance frequency of the word in collocational constructions and the total frequency of the same word. Both measures are included primarily in order to generate expectations for nearest-word neighbours during the SLP parsing and recognition process.

5. Present Status

The syntagmatic information described in this paper is presently incorporated into our Swedish pronunciation dictionary (cf. Huber 1990, Hedelin & Huber 1992). Regarding the inclusion of statistical data concerning the frequency of occurrence of collocations and their constructional tendency, we are of course aware of the three-fold discrepancy:

- that the frequency data provided by Språkdata reflect solely the distribution over the accumulated one-million-words corpus, thus ignoring the dispersion between different text types and genres;
- that these frequencies are based on the evaluation of written news texts, and do not ideally represent the statistical distribution of words in comparable news speech;
- that word frequencies in general require continuous updating both in a macroperspective (i.e. to reflect language changes and the concomitant expansion and reorientation of vocabularies) and in a microperspective (i.e. to capture the prevalence of thematically dictated terminology in application specific domains).

References

Allén, S. (1975) "Frequency Dictionary of Present-Day Swedish based on Newspaper Material, Part 3: Collocations", Almqvist & Wiksell, Stockholm

Fillmore, Ch.J. (1968) "The Case for Case", in: E Bach and R T Harms, Universals in Linguistic Theory, Holt, Rinehart & Winston, Chicago

Hedelin, P. & D. Huber (1990) "The CTH Speech Database: An Integrated Multi-Level Approach", *Speech Communication* Vol. 9, No. 4, pp.365-374

Hedelin, P. & D.Huber (1992) "A Pronunciation Dictionary for Swedish", this Volume

Huber, D. (1989) "Parsing speech for structure and prominence", Proceedings of the First Inter-national Workshop on Parsing Technologies, CMU, Pittsburgh, Penn.

Huber, D. (1990) "An Electronic Dictionary for Computer Speech Applications", Proceedings of the International Workshop on Electronic Dictionaries, Oiso, Japan

Klatt, D.H. (1987) "Review of Text-to-Spech Conversion for English", *Journal of the Acoustical Society of America* Vol. 82, No. 3, pp.737-793

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

Stockwell, R.P., P. Schachter & B.H. Partee (1973) "The Major Syntactic Structures of English", Holt, Rinehart & Winston, New York

Dieter Huber
Department of Information Theory,
Chalmers University of Technology
S-412 96 Gothenburg,
Sweden