A GENERIC LEXICON TOOL FOR WORD MODEL DEFINITION
IN MULTIMODAL APPLICATIONS

Julie Carson-Berndsen
Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
Julie.Berndsen@ucd.ie
http://www.cs.ucd.ie

ABSTRACT
This paper describes a generic lexicon tool which uses lexical representations and finite state transducers enhanced by arithmetic operations in DATR to generate individual output formats from a general phonological feature based representation. The tool was developed in connection with the lexicon component of a diagnostic evaluation toolkit, BEETLE, for a linguistic word recognition system. This lexicon is used online by the system to distinguish between actual and potential syllables and used offline for evaluation purposes with respect to a particular corpus. Rather than design a syllable lexicon which can only be used for these two tasks, it was decided to develop a more generic lexicon from which specific lexica can be generated on the fly; BEETLE is, therefore, only one application of the generic lexicon tool which can be used to generate output formats for other speech technology and multimodal applications.

Keywords: generic lexicon, event-based syllable models, finite state techniques

1. INTRODUCTION
This paper describes a generic lexicon tool which uses event-based syllable models represented in DATR (cf. [5]) and finite state transducers enhanced by arithmetic operations to generate specific output formats from a general phonological feature based representation. The generic lexicon tool has arisen as a further development of the lexicon component of a diagnostic evaluation system, BEETLE, for linguistic word recognition as described by Carson-Berndsen in [4]. The syllable lexicon is used online by the system to distinguish between actual and potential syllables and used offline for evaluation purposes with respect to a particular corpus. Rather than design a syllable lexicon which can be used only for these two tasks, it was more elegant from a computational linguistic point of view to develop a generic lexicon from which specific lexica can be generated on the fly.

Section 2 of this paper introduces the event-based syllable models which are used in the linguistic word recognition system and in the BEETLE toolkit. Section 3 presents the generic lexicon tool and section 4 discusses some other possible output representations which can be generated by the tool. Section 5 concludes the paper and indicates future prospects for this tool in the context of speech technology and multimodal applications.

2. EVENT-BASED SYLLABLE MODELS IN THE LEXICON
The phonological component of the linguistic word recognition system (cf. [4]) requires a multilinear (or autosegmental), temporally annotated representation of syllable structures in the lexicon in order to distinguish between actual structures (i.e. those in the corpus) and potential or out-of-vocabulary structures. The phonological component uses as knowledge, multilinear event-based phonotactic descriptions to determine well-formed structures based on acoustic input. The complete event logic for such multilinear representations in spoken language recognition has been presented by Carson-Berndsen (cf. [4]) building on earlier work by Bird (cf. [1]). The multilinear phonotactic representations cater for coarticulation phenomena and the phonological component of the linguistic word recognition system uses finite state models to implement the underlying event logic theory.

The event-based syllable models define the basis for well-formed word models within the linguistic word recognition system and they will be main focus of this paper. These models consist of multilinear representations of autosegmental tiers of features with their corresponding values and relevant temporal annotations. The synchronisation function across the feature tiers is not uniform, that is to say, the features do not, in general, start and end at the same time. Features are not arbitrarily segmented into phonemic lengths. For example, the German syllable /dOk/ in “Doktor” is assumed to have a single autosegment [voiced] on the phonation tier followed by a single autosegment [voiceless] rather than two successive [voiced] features followed by a single [voiceless] feature. This is the essence of the obligatory contour principle i.e. the stipulation that no two adjacent autosegments have the same value. A stylised graphical representation of the
The diagnostic evaluation toolkit, BEETLE, stipulates a two-stage evaluation terms of logical evaluation with respect to a data model and empirical evaluation with respect to real signal tokens (cf. [4]). In fact Figure 1 is exactly the stylised representation of the syllable model for /dOk/ which can be generated with respect to a data model as is required for the logical evaluation stage. The original lexicon for this diagnostic evaluation was calculated “top-down” on the basis of phonemically labelled signal files using context information to construct average durations of the phone segments and smoothing over features in line with the obligatory contour principle. The resulting structures in the lexicon were event-based multilinear syllable models with temporal annotations. With syllable models of this type, it was possible to access information regarding individual tiers and their melodies.

However, for the linguistic word recognition system, it is necessary to access not only information regarding individual tiers and their structure but also to see how these tiers are related to one another in terms of overlap relations. This type of information is required for the lexicon component and for the generation of reference files for the evaluation procedure. The architecture of the BEETLE toolkit incorporating the generic lexicon tool is depicted in Figure 2. Although the actual output representation format of this information is specific to the linguistic word recognition application, the structure of the syllable models can provide valuable information for both segmental and non-segmental (multilinear) speech technology applications. This provided the main motivation for the generic lexicon tool presented in the next section.

3. THE GENERIC LEXICON TOOL

The generic lexicon tool uses event-based syllable models which are represented in DATR (cf. [5]), a simple language designed specifically for lexical knowledge representation which allows the definition of nonmonotonic inheritance networks with path/value equations. The DATR representation of the event-based syllable models required for BEETLE has been discussed in connection with non-segmental phonology by Cahill, Carson-Berndsen & Gazdar in [2]. The generic lexicon tool was developed on the basis of this description and the event-based multilinear representations were calculated directly in the lexicon using finite state techniques and arithmetic operations.

Within the lexicon, finite state transducers are used to represent event-based syllable models and to represent the linear structure of individual tiers. The lexical representation in DATR consists of the following:
1. Feature templates for vowels and consonants
2. A syllable structure template defining the canonical structure for a particular language
3. Specific syllable entries
4. Specific phonological segment definitions
5. Finite state transducers for performing
   a) Smoothing or spreading (in line with the obligatory contour principle)
   b) Mapping to other output formats.

Given an input query, the generic lexicon tool generates the output format for BEETLE on the basis of the
specific syllable entries by inheriting feature information and average durations for each phonological segment via the syllable structure templates, and the finite state transducers, enhanced by arithmetic operations, perform spreading in line with the obligatory contour principle. A visualisation of the generic lexicon tool is depicted in Figure 3. Rather than representing the somewhat arbitrary splitting of events necessitated by segmentation into phonemic segments, it is possible with this type of representation to refer to a single event which spreads (or is smoothed) over more than one phonemic segment.

In the German syllable /mIt/ “with”, for example, the phonation tier consists of two autosegments [voiced] followed by [voiceless]. The lexical entries for specific phonological segment definitions contain not only feature descriptions, but also average durations for these segments so that it is possible to calculate the example durations for each of these segments directly using the arithmetic enhancements of the Zdatr implementation (cf. [6]). From the general lexicon representation, it is possible to generate the following information about the syllable /mIt/:

1. Tier melody information (here phonation tier as example) with a precedence relation (represented by < ) and temporal annotations (in ms):
   e([voiced], 142) < e([voiceless], 79)

2. Overlap information (for /m/ where the ° represents the overlap relation):
   e([voiced]) ° e([nasal]) ° e([labial])

The temporal annotation in the tier melody has been calculated by a finite state transducer in DATR which models the obligatory contour principle by converting, in this case, two voiced segments with their individual average durations in the phonemic segments /m/ and /l/ to a single autosegment [voiced] with an average duration equal to the sum of the individual durations; this is the event e([voiced], 142).

The overlap information, represented in this way, is interpreted by BEETLE to mean that there exists overlap relations between the events e([voiced]) and e([nasal]), between (e([nasal]) and e([labial]) and between e([voiced]) and e([labial]), although in the axioms of the event logic, of course, overlap is not a transitive relation. This format is thus a convenient abbreviatory device for BEETLE.

This representation of a syllable, in terms of temporal precedence and overlap relations, is used by the BEETLE toolkit to evaluate the linguistic word recognition system. The feature system used here is a multi-valued classification based on the IPA which is required for this specific application. However, the generic lexicon tool has been extended to a feature geometry model (cf. [3]).

4. OTHER REPRESENTATIONS

The main motivation for the development of the generic lexicon tool was to be able to generate many different output representations on the fly. On the basis of a phonological feature description of syllable structure, individual lexica with varying degrees of granularity (e.g. features, phonemes, syllables, etc.) for recognition and synthesis can be generated in an application-specific format. The generic lexicon tool uses either individual finite state transducers or cascades of finite state transducers to generate other segmental and multilinear (nonsegmental) representations from the DATR event-based syllable models.

Some of the other output representations generated by the generic lexicon tool are phonemic representations with and without temporal information, segmental feature representations also with or without temporal annotations, and graph visualisations of the phonological structures of the syllable models. A detailed discussion of the graph visualisations can be found in [3].

Multilinear output representations can consist either of individual tier melody representations, such as the example in section 3, or complete phonological event structures with parallel information describing temporal relations between all features in the representation. Some examples of other multilinear phonological representations such as the encoding of multi-tape finite state transducers as proposed by Wiebe (cf. [8]) or the mapping to the feature set proposed in Kirchoff’s delayed synchronisation approach to speech recognition (cf. [7]) are described by Cahill, Carson-Berndsen & Gazdar in [2].

5. CONCLUSION AND FUTURE PROSPECTS

This paper presented a generic lexicon tool which uses event-based syllable models represented in DATR and individual or cascades of finite state transducers enhanced by arithmetic operations to generate the individual output formats from a general phonological feature based representation.

In addition to the particular application of the generic lexicon tool as part of the diagnostic evaluation toolkit,
BEETLE, the event-based syllable models provide information for word model definition in speech technology and multimodal applications.

The speech technology applications concern the generation of application-specific lexica of varying granularity for speech recognition and speech synthesis systems and an example of a particular format for the phonological component of a linguistic word recognition system was provided in section 3. Since the information is represented in the generic lexicon as multilinear (parallel) structures, information regarding the temporal coordination of these structures in terms of overlap and precedence relations may be generated.

In the context of this work, ‘multimodal’ refers to the output of parallel information which is structured with respect to temporal and spatial relations and thus the multilinear representations in the lexicon can be enhanced by further features to provide the basis for the generation of such structured information. This type of structure is of paramount importance in multimodal applications where speech and gesture have to be coordinated.

The generic lexicon representation in DATR is a powerful tool for generating syllable lexica for specific purposes and for providing structured and coordinated information for multimodal applications.

6. REFERENCES


