TALN 2004, Session Poster, Fès, 19–21 avril 2004

Multiple Lexicon Generation based on Phonological Feature Trees

Moritz Neugebauer, Stephen Wilson
Department of Computer Science - University College Dublin
Belfield, Dublin 4, Ireland
moritz.neugebauer;stephen.m.wilson@ucd.ie

Résumé - Abstract

De manière générale, les linguistes informaticiens utilisent les structures de données arborescentes pour la documentation et l'analyse des données morphologiques et syntactiques. Dans cet article nous appliquons de telles structures sur des données phonologiques et nous démontrons comment de telles représentations peuvent avoir des applications utiles et pratiques en lexicographie informatique. À cet effet, nous décrivons trois modules intégrés: Le premier module définit un ensemble de caractéristiques multilangages dans une structure arborescente exprimée en XML; le deuxième module parcours cet arbre et établis une généralisation sur des données contenues dans cet arborescence, optimise les données phonologiques et mets en valeur les implications des caractéristiques. Le troisième module utilise l'information contenue dans l'arborescence comme une base de connaissance pour la génération de syllabes lexiques à caractéristiques multiples.

Tree-based data structures are commonly used by computational linguists for the documentation and analysis of morphological and syntactic data. In this paper we apply such structures to phonological data and demonstrate how such representations can have practical and beneficial applications in computational lexicography. To this end, we describe three integrated modules: the first defines a multilingual feature set within a tree-based structure using XML; the second module traverses this tree and generalises over the data contained within it, optimising the phonological data and highlighting feature implications. The third uses the information contained within the tree representation as a knowledge base for the generation of multiple feature-based syllable lexica.

Mots-clefs – Keywords

Lexicographie, Représentations phonologiques, XML
Lexicography, Phonological representations, XML

1 Introduction

The lack of a standardised scheme for the annotation of phonological information is in sharp contrast to other research areas within computational linguistics and natural language processing where numerous schemes exist. A comprehensive collection of tools and formats for creating and managing linguistic annotations are provided online by the Linguistic Data Consortium. We aim to aid research into standards of this kind by presenting a structured format for the encoding of phonological information. The principal motivation for our work was twofold: a. the development and use of consistent and coherent encoding formats for data representation, as well as standardised schemes for annotation of linguistic information and b. the development of reusable, integrated systems and tool architectures for language processing and analysis, including the corresponding development of a data architecture to best suit research needs (Ide, 1999).
Although the encoding format presented lends itself easily to use within a wide range of NLP research fields, this paper focuses on its application in computational lexicography, specifically to the generation of multiple phonological lexica. The Extensible Markup Language (XML) is used frequently to model natural language data in large scale applications and is employed here to define the structure of phonological feature trees (Neugebauer & Wilson, 2004). The parsing of a marked up document in order to retrieve the data contained within it is a common application task to which the Document Object Model (DOM) is key as it defines a set of interfaces for referring to, retrieving and changing items within an annotated structure. In the following sections, we outline the process that defines such phonological tree structures, describe the feature optimisation module and explain how our lexical generation mechanism exploits the phonological information contained within our encoding format to create multiple lexica from a single foundation lexicon.

2 Phonological Feature Trees - Acquisition

Phonological feature trees are data structures that define a set of user specified symbol-to-phonological feature attribute mappings. They are repositories of phonological information explicitly linked to particular symbol sets. Their creation was driven by the need to have a structured symbol-feature inventory that could be used as a knowledge source for phonological document generation and mapping. The acquisition module described here, outlines how user-driven definition of such repositories annotates and stores the data within a useful and coherent data structure.
The acquisition module provides a graphical user environment for the annotation of these mappings and stores them in a structured XML-tree called a *feature profile*. Feature profiles are multilingual resources as they are intended to define a full inventory of phonological feature information for a complete symbol set across a number of languages. Thus, from an abstract feature set, language-specific symbol-to-feature mappings are constructed.
The data structure described here not only encodes associations between symbols and one particular feature set, it is intended to annotate mappings between a symbol set and multiple feature sets. Thus phonological feature trees may model associations between – for example – the SAMPA notation set and phonological features inspired by the IPA, binary features, and even features from other modalities (e.g. mappings between symbols and visual features). It is this ability to encode mappings across numerous feature sets that enables phonological feature trees

to act as the knowledge base for multiple lexicon generation.
Phonological feature trees as discussed in this paper form the second and third phases of the production process shown in Figure 1. A finite-state automaton modelling the legal combinations of sounds for a supplied domain – a phonotactic automaton – is partially learned (phase 1); this structure is then used to automatically generate the interface for the feature definition module (phase 2); once feature-to-symbol associations have been created in phase 2, the optimisation phase generalises over the feature set, supplying additional information regarding logical relations among individual as well as sets of features (phase 3); finally the information contained within the phonological feature trees is used by the lexical generation module (phase 4). As indicated in the top box in Figure 1 we require linguistic annotation at least at the segment level (depicted by "$\geq$ segment level").

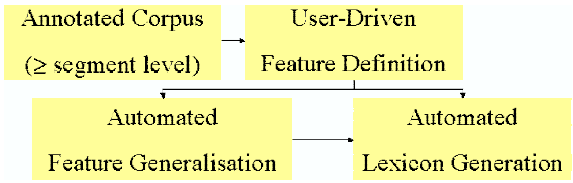

Figure 1: Production process

A training set of syllables is input into a grammar induction tool to learn the structure of a deterministic stochastic automaton from a set of training sequences (Carson-Berndsen & Kelly, 2004). A finite-state machine is induced which models all legal combination of sounds within the syllable domain based on the training data. This finite-state machine serves as the segment-annotated corpus in our example. The feature definition module takes this phonotactic automaton as its input and extracts every unique occurrence of a phonological symbol from the automaton and dynamically creates a graphical user environment that allows the user to define feature associations for those symbols. By automatically creating the interface, the module ensures that users define associations only for those symbols that occur in the data set.
The interfaces of the definition module provide for numerous means of feature annotation (e.g. different arity of features).
From this data a Document Type Definition (DTD) is automatically generated. This DTD will be used to provide top-down constraints in the creation of subsequent profiles that use the same feature set. A section from a DTD describing a multi-valued feature profile is shown below, where an attribute like *manner* might take values describing broad sound classes.

```
<!ELEMENT featureProfile (featureAssociations)*>
<!ELEMENT featureAssociations (symbol,features*)>
<!ELEMENT symbol (#PCDATA)>
<!ATTLIST symbol notation ( IPA | SAMPA) #IMPLIED>
<!ELEMENT features (phonation?,manner?)>
<!ELEMENT phonation (#PCDATA)>
<!ELEMENT manner (#PCDATA)>
<!ELEMENT place (#PCDATA)>
```

Figure 2: Document Type Definition

Once the full feature set has been specified, the module generates a graphical representation of the data set (Figure 3). Users then simply create associations by pointing and clicking, first at the symbol and then at the features that are to be associated with it. Feature associations are added to the feature profile upon user confirmation.

Figure 3: User Interface Generation (selection)

The symbols used have an underlying IPA-Unicode representation, however a notation transducer within the module allows for the mapping between a number of notation conventions (e.g. SAMPA, Wordbet, ARPAbet etc.). Any such mappings maintain the structural integrity of the phonological feature tree, namely symbols are mapped to their corresponding cognates as defined by the transducer, user defined feature associations are mapped to themselves. To summarise, the module provides a number of graphical interfaces for the following procedures: a. the graphical display of a phonological feature tree and all its nodes containing articulatory information as well as the language for which it is defined, b. graphical editing of the information contained within the trees – addition and deletion of nodes/tiers, modification of articulatory data. Any changes that have implications for DTDs (e.g. addition of an extra tier of information) are implemented automatically and c. selection of particular functions for the manipulation of the data contained within feature trees, e.g. the extraction of a language specific profile from the superset of all feature associations.
Other interfaces include the notation transducer already mentioned and an interface to the lexical generation mechanism described below. The following section describes how we can determine generalisations about the information contained within a feature profile.

3 Phonological Feature Trees - Generalisation

The above motivations for defining phonological feature profiles lead to an expressive knowledge base which provides a fine-grained level of description for the modelling of individual phonological segments. However such a rich set of features, while having extensive descriptive value, is not particularly suited to user-driven manipulation such as the identification of implicational relations between (sets of) features. Our goal is to obtain this valuable information while limiting the need for manual effort and to this end we propose a computational method based on automated deduction which delivers correspondences between individual features as well as all sets of sounds created by combinations of those. Annotated segement entries as defined by phase two of the production process (Figure 1) represent the input for phase three as we seek to automatically extract information about feature distribution within our database.
The optimisation module traverses the phonological feature tree, applying our algorithm with a view to performing as much deterministic inference as possible. In this way, we automatically generate feature hierarchies, similar to information hierarchies in unification-based grammar formalisms, ordering features with respect to the size of their extents, i.e. the segment sets they describe.[1] This information enriches the phonological feature profiles with two elements that

[1]The denotational semantics of XML do not provide for multiple inheritance, therefore, we choose to "multiply out" every single combination of features to achieve its extent in terms of phonological segments.

distinguish between those features which imply other features and those which do not (Neugebauer & Wilson, 2004). If the symbol-to-feature mappings are manipulated, updates to our knowledge base are carried out using XSL, the stylesheet language for transforming XML documents.

4 Lexicon Generation

The acquisition of the featural data stored in phonological feature trees as described above is an incremental process. Over time, phonological information from a particular feature set is associated with a complete set of symbols. Given a set of alternative features, instead of creating a separate feature tree for that particular phonological data set, it seems a more efficient use of the information to store them within a superstructure that models symbol-feature associations over a number of feature sets. By explicitly linking symbols with a number of associations of phonological features, we can facilitate the smooth mapping of symbols from one feature set into another. It is this process that forms the basis for multiple lexicon generation.
Given linguistic documents that are based on specific phonological features, it may at times prove useful to generate additional documents that are structurally identical to the originals but that use phonological information from alternative feature sets. In doing so, speech scientists can examine the suitability of various feature sets with respect to certain applications. An example might be the suitability of different feature sets for parametric speech synthesis. In this section, we outline how phonological feature trees can be used to generate multiple feature-based lexica that are structural clones of each other but which source their phonological information from different feature sets. The lexica that are discussed here are feature-based syllable lexica represented in XML. The foundation lexicon is automatically created by LeXMLicon, a lexical generation mechanism (Wilson *et al.*, 2003). Lexica created by this mechanism not only model the featural information for a segment, but also its position within the syllable. An example entry is shown below.

```
<syllable>
 <lexeme>ha</lexeme>
  <onset type="first">
   <segment phonation="voiceless" manner="fricative" place="glottal"
    duration="175">h</segment></onset>
  <nucleus type="first">
   <segment phonation="voiced" manner="vowellike" place="back"
    height="low" roundness="nonround" length="short" duration="232">a
   </segment></nucleus>
</syllable>
```

Figure 4: Sample entry from foundation lexicon

The feature set of the foundation lexicon is based on the sound descriptions provided in the IPA. Similar phonological representations of lexical items have been proposed ((Tiberius & Evans, 2000), (Cahill *et al.*, 2000)), although in their work DATR is used to model the data. LeXMLicon, while using the inference mechanisms of DATR to construct a concise phonological lexicon, uses XML to model the output data (Figure 4). By grouping a number of these feature trees into a super-tree that models feature associations from different phonological feature sets with particular symbols, the syllable structures that contain those symbols can map easily between feature sets.

5 Conclusion

The generation of multiple lexica from one foundation lexicon makes use of two interacting mechanisms: one for the creation of a foundation lexicon and another that oversees the mapping of this foundation lexicon onto other additional lexica with different phonological information. The output is an XML representation of the syllable, that models segmental units as well as their syllabic position. This constitutes the foundation lexicon from which all others will be generated. During the subsequent generation of additional lexica, all syllabic forms contained within the foundation lexicon are mapped to identical structures in the new lexicon. However, by using the feature tree as a knowledge source, we use previously defined feature associations encoded within it to model the segmental entries within the new lexicon. In this way, multiple lexica based on different feature sets can be created.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. 02/IN1/ I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

Références

CAHILL L., CARSON-BERNDSEN J. & GAZDAR G. (2000). Phonologically based lexical knowledge representation. In F. V. EYNDE & D. GIBBON, Eds., *Lexicon Development for Speech and Natural Language Processing*, p. 77–114. Dordrecht: Kluwer.

CARSON-BERNDSEN J. & KELLY R. (2004). Automatic induction of multilingual phonotactic resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

IDE N. (1999). Encoding linguistic corpora. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

NEUGEBAUER M. & WILSON S. (2004). Phonological treebanks – issues in generation and application. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon.

TIBERIUS C. & EVANS R. (2000). Phonological feature based multilingual lexical description. In *Proceedings of Traitement automatique des langues naturelles*, Lausanne.

WILSON S., CARSON-BERNDSEN J. & WALSH M. (2003). Enhancing phonological representations for multilingual speech technology. In *Proceedings of the International Congress of Phonetic Sciences*, Barcelona.