Phonological Treebanks
—
Issues in Generation and Application

Moritz Neugebauer, Stephen Wilson

Department of Computer Science
University College Dublin, Ireland
{moritz.neugebauer, stephen.m.wilson}@ucd.ie

Abstract

The continuing popularity of XML as a data exchange format and the concurrent rise of treebanks as natural language resources within various domains of language processing have led us to extend their domain of application to phonological data. Typically, treebanks are a language resource that provides annotations of natural languages at various levels of structure and in this paper we present a tree-based format to capture phonological information at the syllable level, at the segment level and even including more fine-grained featural information. Two integrated modules in relation to phonological treebanks are described: the first uses a multilingual feature set to augment segment-annotated corpora in terms of a tree-based structure represented in XML. The second module allows these feature trees to be traversed and the data contained in it to be optimised in a purely data-driven manner.

1. Introduction

Over the past few years, treebanks have become important in various areas of computational linguistics ranging from syntactic to morphological and semantic-pragmatic applications. The continuing popularity of XML as a data exchange format and the concurrent rise of treebanks as natural language resources within the above domains have naturally led us to extend their domain of application to phonological data. Typically, treebanks are a language resource that provides annotations of natural languages at various levels of structure and in this paper we present a tree-based format to capture phonological information at the syllable level and at the segment level including articulatory information.

Two integrated modules in relation to phonological treebanks are described: the first defines a multilingual feature set within a tree-based structure represented in XML, the second traverses this feature tree and optimises the data contained in it, highlighting feature redundancies. The mapping component of the integrated system then takes the information contained within the feature tree and uses it to augment a specific phonological representation, the Multilingual Time Map (Aioanei et al., 2004). The integrated system described here takes the form of a graphical user environment, which presumes no knowledge of the technologies used on the part of the user.

2. Segment-Based Phonotactic Descriptions

A phonotactic automaton is a finite state machine that models all permissable combinations of sounds for a language within the syllable domain. A Multilingual Time Map extends this model, being conceptually an n-tape finite state transducer that models not only the phonotactics of a language, but also a wide (potentially limitless) range of additional segment-specific linguistic knowledge, e.g. the corpus frequency of a segment, its graphemic equivalent and so on (Carson-Berndsen, 2002). Specifically, this paper is concerned the augmentation of the multilingual time map with data tapes relating to a segment's associated phonological features, as well as tapes that highlight feature implications for a given feature set. This paper focuses on the final two stages of a three part production process, namely the feature definition stage and the feature optimisation stage. Although space prohibits an in-depth discussion of stage one – inducing phonotactics – a brief outline is necessary. The initial stage of production assumes the existence of a segment labelled syllable corpus for a language. Taking this data as input, the induction module infers a phonotactics for the supplied corpus. In addition, it induces certain statistical properties regarding segments that are corpus specific, including weight and probability. The output from the module is an XML representation of a finite state transducer – in essence a multilingual time map – that has three data tapes: phoneme, weight and probability. The following sections concern the acquisition of a secondary data representation, a phonological feature tree, that will be used to augment the multilingual time map with the additional tapes mentioned above.

The creation of phonological treebanks was motivated by the desire to create structured repositories of phonological information explicitly linking feature sets with symbol sets and which could be sourced for use during document generation and mapping. The module described here provides an environment for user driven acquisition of such repositories, annotating and storing the information in a useful and coherent data structure – phonological feature trees. Phonological feature trees are intended to be multilingual resources in the sense that they should provide a full inventory of phonological features for a complete symbol set across a number of languages. The process of acquiring this inventory is designed to be incremental: each defined set of symbol-to-feature associations is annotated with respect to a particular language, indicating the language(s) for which that particular combination of features is valid. Smaller subsets of associations, which can be considered as individual profiles for particular languages, can naturally be extracted as required.

The advantages of using XML as a data exchange format are many. Among them is the provision of a standardised encoding format which allows ready access to data via a wide range of programming APIs. Within the context of wider research goals concerning the portability of linguistic resources, XML ensures that the data contained within phonological treebanks is reusable by a wide group of people and applications, as any user-required idiosyncratic structures and program-specific formats can be easily generated.

The phonological trees discussed in this paper are used in the augmentation and management of a number of specific phonological representations. They form the second and third phases of the cycle shown below describing the production of a multilingual time map. We seek to partially learn the time map's structure in phase 1 of the above diagram; this structure is then used to 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, adding additional information regarding logical relations among individual as well as sets of features (phase 3). The diagram below shows all three phases, while in this paper we focus on the *Feature Definition* and the *Feature Generalisation* components pictured below.

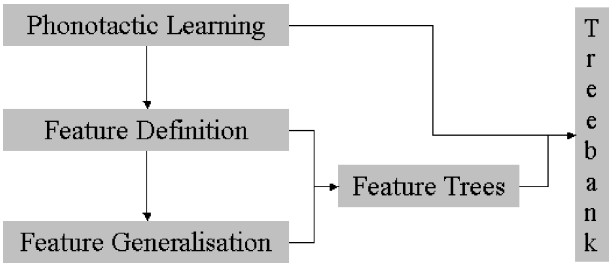

Figure 1: Visualisation of the annotation process

The cycle's initial phase sees a set of syllables input to an in-house tool named *Phonotactic Automaton Learner* (PAL) which automatically generates a finite state machine that models all the legal combinations of sounds within the supplied domain (Carson-Berndsen and Kelly, 2004). This structure forms the basis for the multilingual time map transducer. We wish to augment each transition within the time map with an additional tape that supplies information about phonological feature overlap relations and the feature definition module provides the means to do so (Carson-Berndsen and Kelly, 2004).

3. Phonological Treebanks: Feature Set Definition

A feature definition module facilitates the user driven acquisition of a multilingual feature set. The module provides for the specification of user defined symbol to feature mappings at varying levels of granularity within an intuitive graphical environment. Features may be defined as unary, binary or multi-valued entities as described below and are intended to be multilingual in the sense that a full feature inventory is defined for a particular symbol set across a number of languages. Thus, from an abstract feature set, language-specific symbol-to-feature mappings are constructed. These mappings as defined by the user, are stored in a coherent data structure, a phonological feature tree, which is represented in XML and which forms the basic structure from which fully specified phonological treebanks are created.

Symbol-to-feature attribute mapping describes the explicit linking of a particular phonological symbol with a number of phonological features. Thus, one such association might define the mapping between the initial symbol in [l eh t] (ARPABET transcription for English *let*) and the features *nonvocalic, coronal, consonantal* and *lateral*. The module combines data driven approaches to interface generation with user-driven data definition. Features can be encoded as one of three feature structures: unary, binary or multi-level. Unary features can be viewed as properties that can be assigned to segments as stand alone attributes; binary features are attribute-value pairs that have two mutually exclusive values; multi-valued feature structures consist of a number of different data tiers, each of which has an associated set of phonological features as parameters, from which one is chosen. Essentially all three methods of encoding require the input of the full selection of values for a particular feature set. Taking the feature set input, a Document Type Definition (DTD) is automatically generated, which is used to provide validation constraints on subsequent feature trees based on the same feature set. A graphical interface is also generated allowing users to make mappings between symbols and features via point-and-click. It was a deliberate decision to remove all need for users to have any working knowledge of the denotational semantics of XML when encoding feature associations. By using a series of graphical interfaces to define these mappings, we ensure that users can work within an intuitive environment in order to annotate the mappings between symbols and features and allow the module's internal mechanisms to encode these mappings within the structured XML feature tree. The underlying representation of symbols is IPA-Unicode, however, the module's notation transducer allows feature sets that have been associated with this notation to be mapped to a variety of other phonetic alphabets (e.g. SAMPA, ARPAbet). Alternatively, users may explicitly select a notation other than IPA before creating a new feature profile.

The feature definition module also provides interfaces for the graphical display of the structure of the feature tree, showing the articulatory information contained in its nodes, as well as the languages for which it is defined. A second interface provides for the graphical editing of the data contained within the structure, e.g. the addition removal and modification of nodes, tiers etc. Any editing changes that impact on the tree's structure as defined by its DTD - the removal of a tier for example - cause the DTD to be automatically updated subject to user confirmation. The module also has interfaces that allow users to select particular functions for the manipulation of the data within the profile, e.g. extract language specific associations from the superset of all associations; load the structure into an in-house lexical generation mechanism (Wilson et al., 2003) and use it for the generation of multiple lexical documents (Neugebauer

```
<!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: Example: Document Type Definition using multivalued features

and Wilson, 2004).
The design of the module outlined above was motivated by the desire to have tools and interfaces that allowed users to rapidly and easily define a multilingual feature set and link it with a number of symbol sets. The module described, with its emphasis on graphical interaction, (point and click, menu selection etc.), certainly achieves this. Moreover, modelling the data as an XML tree ensures that the annotated data is not tied to any one application or use but can be used and reused by a range of applications for a number of purposes. The data structure provides the basis for the creation of phonological treebanks and the following section describes how we seek to make generalisations over the data, optimising feature set and augment the tree to model this.

4. Induction of Structural Relations among Feature Trees

The segmental annotation of speech data and its enrichment with more detailed information in terms of phonological features is either implicitly or explicitely based on a classificatory system. Such a system potentially displays a number of interesting linguistic generalisations, in this paper for commonalities between segmental units. For example, certain pieces of information may permanently cooccurr in a given set of annotated data. This means that the interaction of phonological labels such as *voiced* and *vocalic* can be captured as a set of implicational rules. In this section we will present an automated means of determining these rules and integrating them into our XML-based format.
With regard to treebanks these generalisations over linguistic annotations address two architectural issues for annotation models as described in (Bird and Liberman, 2001). The first issue concerns cases which involve partial information. For an annotated corpus this means that not all linguistic units have been provided with information on all levels of annotation. Building on a subset of fully annotated data, implication rules allow us to fill possible informational gaps. Partial annotations can consequently be completed based on other annotations which are considered complete. Another issue is that of redundant information which in extension to the cases mentioned in (Bird and Liberman, 2001) also occurs between individual labels on the same level of annotation. For instance, if two feature labels are relevant for one segmental unit only, one of them can be considered redundant. This information can be captured in terms of a bidirectional implication rule as computed by our application.
We will now describe the automated induction of structural relations among phonological features trees like the ones which were generated along the lines of Section 3. Consider the following example of the vowel [ah] which includes a segmental annotation and an annotation in terms of phonological features (cf. Figure 3).

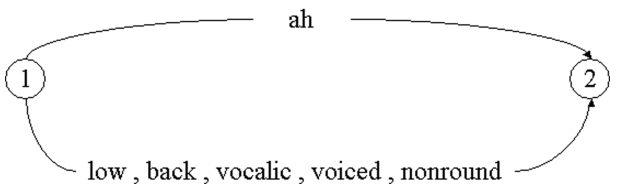

Figure 3: Two-level annotation for [ah]

Since it is the purpose of phonological features to define classes of sounds and each segment carries more than one feature, implicational relations between phonological features are relations between sets of segments. Therefore, our approach is based on the set-theoretical relation of subsumption and the resulting set of rules may be visualised in terms of a subsumption hierarchy. For the above example, we proceed by computing the extents for all phonological features. The comparison of the obtained sets results in a subsumption ordering where unidirectional implications are determined as well as bidirectional ones. The latter case occurs if two features denote the identical set of segments. For a small set of vowels we computed rules of the following kind:

low	→	back , voiced , vocalic , nonround
back	→	–
vocalic	→	voiced
voiced	→	vocalic
nonround	→	–

Figure 4: Examples for induced implication rules

The first example is a unidirectional implication stating that the feature *low* implies the feature *back* among all other features which together fully specify the vowel [ah]. While the feature *back* does not imply any other feature (just as *nonround*), we observe a bidirectional implication between the features indicating phonation and manner of articulation. For the case of this example, it can be said that one of the features (*vocalic* or *voiced*) is redundant since it adds no expressivity to our classificatory system. Next we show how the information which has been induced automatically is integrated with the existing annotations.
Basically, the induction component splits set of features into two subsets: features that are unique for a particular segment and those which are not. Consequently, we in-

troduce two new arcs for each annotation as shown in the figure below (which is an extension of Figure 3).

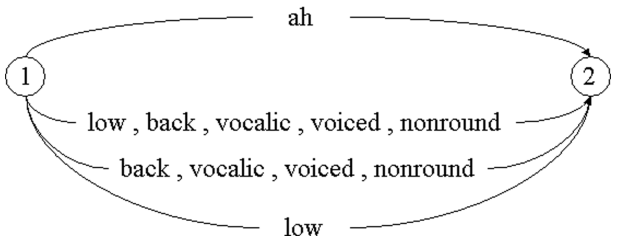

Figure 5: Four-level annotation for [ah]

As expected, the two new arcs together contain all elements present on the feature arc. The fact that *low* appears on arc indicates that based on this information all other specifications can be inferred. If a segment specification does not allow a unique feature to be singled out, the bottom arc would be empty and all features would appear as shared features (just as *back , voiced , vocalic , nonround* in Figure 5). The bidirectional implication between *vocalic* and *voiced* is not expressed in Figure 5. Since this particular rule is not specific to one segmental unit (such as [ah]), but rather holds for the set of all vowels, the rule is captured elsewhere in our database.

The automatic induction of feature implication as described above is fully automated which becomes important when a rich set of symbol-to-feature mappings is initially defined and then frequently manipulated. Our generalisation component provides an efficient handling of this task in the domain of phonological treebanks.

5. Conclusion

Our approach presents a novel application of treebanks to phonological data and this paper focussed on the technology for their generation and application. This finally yields a fine-grained characterisation of segmental units in terms of user-defined symbol-to-feature mappings. Generalisations over this set of all lexical entries allow us to split the set of characteristic features for each segment into shared ones and features which are unique for a specific phonological segment. By these means, even a fairly large multilingual feature set can be maintained in a treebank as well as mined for language-dependent and language-independent phonological implications.

The use of XML ensures that all annotations along the format of phonological feature trees are intelligible across applications. Additionally, the presented annotation format is expressive enough to accommodate user-defined preferences regarding the formal specifications in the phonological feature system. In the last section we gave an example as to how the automatic extraction of linguistically interesting patterns can be facilitated.

Finally, the phonological treebanks have been presented as easily maintainable multi-level knowledge bases of feature- as well as segment- and syllable-labelled data. Future research is primarily concerned with the generation of treebanks for various languages. Another topic will be the integration of even more levels of annotations such as acoustic, allophonic and timing information.

6. 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.

7. References

Aioanei, Daniel, Julie Carson-Berndsen, Anja Geumann, Robert Kelly, Moritz Neugebauer, and Stephen Wilson, 2004. A multilingual phonological resource toolkit for ubiquitous speech technology. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon.

Bird, Steven and Mark Liberman, 2001. A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

Carson-Berndsen, Julie, 2002. Multilingual time maps – portable phonotactic models for speech technology applications. In *Proceedings of the LREC 2002 Workshop on Portability Issues in Human Language Technology*. Las Palmas.

Carson-Berndsen, Julie and Robert Kelly, 2004. Automatic induction of multilingual phonotactic resources. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon.

Neugebauer, Moritz and Stephen Wilson, 2004. Multiple lexicon generation based on phonological feature trees. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*. Fez.

Wilson, Stephen, Julie Carson-Berndsen, and Michael Walsh, 2003. Enhancing phonological representations for multilingual speech technology. In *Proceedings of the International Congress of Phonetic Sciences*. Barcelona.