Phonological feature based multilingual lexical description

Carole Tiberius and Roger Evans
Information Technology Research Institute
University of Brighton
Brighton, UK
{Carole.Tiberius,Roger.Evans}@itri.brighton.ac.uk

Abstract

This paper presents a framework for compactly describing word forms in terms of phonological features. Using a highly modular default-inheritance based approach, the framework supports the description of lexical generalisations traditionally modelled as morphology and phonology in a single phonology-based representation. This representation is more uniform and more detailed than previous approaches of this kind, allowing us to capture generalisations within a language and between related language elegantly and flexibly. The framework is illustrated with examples taken from English, German, Dutch and Danish.

1. Introduction

This paper presents a framework for compactly describing word forms in terms of phonological features. The principal components of a lexicon in this framework are definitions which correspond to *lexemes* – families of words related by inflectional or derivational processes. Given such a lexeme definition, it is possible to read off the phonological structures associated with the word forms.

We restrict ourselves to a segmental representation of the phonology, describing each segment using a set of features, such as *phonation*, *manner*, *roundness*, and *length*. Internally, generalisations over words are captured using default inheritance. Our description framework applies to words within one language, but also across different languages, allowing high level cross-linguistic generalisations to be captured.

The framework covers aspects of word formation that traditionally encompass both morphology and phonology. However, there is no overt representation of morphology here: all the generalisations made are generalisations over phonological structure. In this sense the approach is *data-driven*, where the 'data' consists of the phonological structure of words and all generalisations are ultimately about some aspect of that structure. The only hint of morphology is in the operators associating word forms to lexemes, which have morphosyntactic names such as *plural* and *definite*, but even these relate structure which is in fact phonological.

The framework is implemented in the lexical description language DATR (Evans & Gazdar,

1996). This provides both a formally rigorous foundation for the lexicon and a computational implementation with which to test the definitions, and from which to compile runtime lexicons in appropriate formats for different NL applications. The intention is that a framework like this provides a high level, highly structured, easily maintained and extended lexical resource. Although it can be utilised directly as a computational lexicon (using a DATR query engine), high performance NL applications will typically use representations automatically derived from the DATR code, and possibly highly optimised for the particular application in hand.

This work is perhaps most straightforwardly viewed as a development of the PolyLex lexical description framework (Cahill & Gazdar, 1997; Cahill & Gazdar, 1999a; Cahill & Gazdar, 1999b). PolyLex adopts a similar formal and theoretical approach to lexical description, but only describes words to the level of phonological segments. In addition it does include some explicit morphological machinery. Our proposal extends the PolyLex word model down to the level of phonological features, and adopts a more uniform phonologically-based approach to lexical generalisation. This allows us to capture more generalisations more precisely, and to organise our lexicon into distinct self-contained modules corresponding to levels of lexical description (lexeme, syllable, segment etc.)

This paper is organised as follows. Section 2 discusses the theoretical background and related work in this area. Section 3 and 4 present the framework itself: the way word forms are represented, and how hierarchical descriptions of word forms are organised. In section 5, we illustrate the framework with examples using Dutch, English, German and Danish nouns. Section 6 outlines current areas of development.

2. Theoretical background

The traditional view of word formation treats morphology and phonology as distinct aspects of a word's internal structure. However there is strong evidence that these two aspects interact in the word formation process. For example allomorphic variation of affixes is frequently determined by phonological context and affixation itself often imposes phonological requirements. There are also nonaffixational morphological relations, such as umlaut, whose origins are purely phonological. Such phenomena make a clean definition of a morphology/phonology interface problematic. In broad terms, two main positions on this issue can be distinguished in the literature: *noninteractionism* in which morphology strictly precedes phonology, providing abstract structures on which phonological rules operate (e.g. SPE model (Chomsky & Halle, 1968)), and *interactionism* which allows some phonological operations to precede morphological ones (e.g. Lexical Phonology (Mohanan, 1986)).

In the last decade, theories have been developed which take an intermediate position, notably Cahill (1990; 1993) and Gibbon (1992). The idea behind Cahill's *syllable-based morphology* is that since many morphological alternations are phonologically based, they can be best described as mappings between sequences of tree-structured syllables. Effectively, morphological operations are defined in terms of changes to the phonology. For example, German umlaut (Apfel — Äpfel) will be represented as a change of the vowel (peak) of the first syllable. Our approach draws heavily on the theoretical work from Cahill (1990) which has been further developed in Cahill (1993) and Cahill and Gazdar (1997; 1999a). Gibbon (1992) adopts a very similar position, although his work is more tuned towards lexicons for speech applications integrating phonological information above the level of the syllable, such as metrical structure, whereas the focus here is the structure within syllables.

Both these approaches still make a distinction between morphology and phonology. For example, Cahill and Gazdar (1997; 1999a) define the morphological form of a word in terms of the phonological form of its root and the morphological form of a suffix. However, they do not adopt the traditional notion of level of description, or of rules mapping from one level to the other. The linguistic description is just a set of simultaneously applicable constraints. These constraints may, for example, directly connect morphosyntactic attributes to individual phonological components of word forms.

Our work pushes this view further making no sharp distinction between morphology and phonology at all. We start with the actual structures that we are aiming to describe, and generalise over them motivated purely by structural considerations, without any preconceptions about whether the generalisations are phonological or morphological. Our initial structures are phonological, and so our generalisations are phonological. There are echos of traditional morphology, for example generalisations which correspond to some extent to traditional 'morphemes', but no explicit separation or reference to morphology is required. To put it another way, we view morphology as 'just' abstractions over phonology, and by utilising a sufficiently powerful abstraction language to describe the phonology, we obviate the need for a separate morphological language.

Nevertheless, our approach does reveal *structural* distinctions which induce a high degree of modularity in our representations. Like Zwicky (1990), the framework explored here opts for something like the subcomponent divisions of traditional grammar, rather than the level or strata of "lexical morphology/phonology". Thus we distinguish lexemes, syllable sequences, syllables and phonemes, so that even a word consisting of just one phoneme requires corresponding syllable, syllable sequence and lexeme structures to be defined.

Following Cahill and Gazdar (1997; 1999a), we adopt a segmental model of phonology in which phonological units are discrete and in simple temporal sequence. However, rather than using phonemic transcriptions, where the primitives are vowels and consonants, we go down to the level of phonological features. This permits a more accurate and a more elegant treatment of phenomena such as elision, final consonant devoicing, vowel lengthening, and assimilation (e.g. Cahill (1993), Coleman (1992), Bird and Klein (1990)). For example, vowel lengthening involves just a change in the length feature of the vowel, regardless of the particular vowel involved, whilst final consonant devoicing just changes the voice feature of final consonants[1].

We concentrate on the treatment of inflection and in particular the syllable structure of inflected forms. We do not currently address higher level issues such as metrical structure or lexical stress. Our general approach is of the *inferential-realisational* type (e.g. Zwicky (1985; 1990), Anderson (1988), Stump (1993a; 1993b)). In theories of this type, paradigms (inflectional classes, declensions, conjugations, etc.) are treated as analytically central, rather than epiphenomenal or of secondary status. The central notion in these theories is the lexeme, not the word or the morpheme. Words exist as realisations of morphosyntactic specifications of lexemes: an inflected word's association with a particular set of morphosyntactic properties licenses the application of rules determining the word's inflectional form. For example, the English word *likes* arises by means of a rule appending *-s* to any verb stem associated with the properties "third-person singular subject agreement", "present tense", and "indicative mood". In our framework lexemes, represented as DATR nodes, are the primary content of a lexicon and word forms are accessed by applying lexical operations (implemented using the lexical rule

[1]Of course, a segmental description is still an idealisation of reality – see Cahill, Carson-Berndsen, and Gazdar (2000) for a discussion of how a segmental description can be extended to deal with nonsegmental issues.

techniques described in Evans, Gazdar and Weir (1995; 2000) and Smets and Evans (1998)) such as *singular*, *definite* or *third-person* to lexemes.

Our framework makes extensive use of default inheritance to capture linguistic generalisations. In this sense it is closely related to Corbett and Fraser's Network Morphology (Corbett & Fraser, 1993), which treats language as a network of interacting parallel hierarchies of linguistic knowledge. However, in Network Morphology the hierarchical structure is motivated by theoretical and typological principles, in contrast to our more data-driven view. For example, we might introduce the notion of a suffix, not because we are following a linguistic theory that presupposes a suffixation operation, but because a suffix captures a generalisation about final components of many word forms.

3. The structure of word forms

The primary objects our lexicon aims to describe are word forms, or more precisely phonological analyses of word forms. We view such word forms as labelled tree structures with a root representing the whole word form and successive decomposition into syllable sequences and then syllables. An example of the tree structure associated with the singular of the lexeme *Hand*, is given in figure 1.

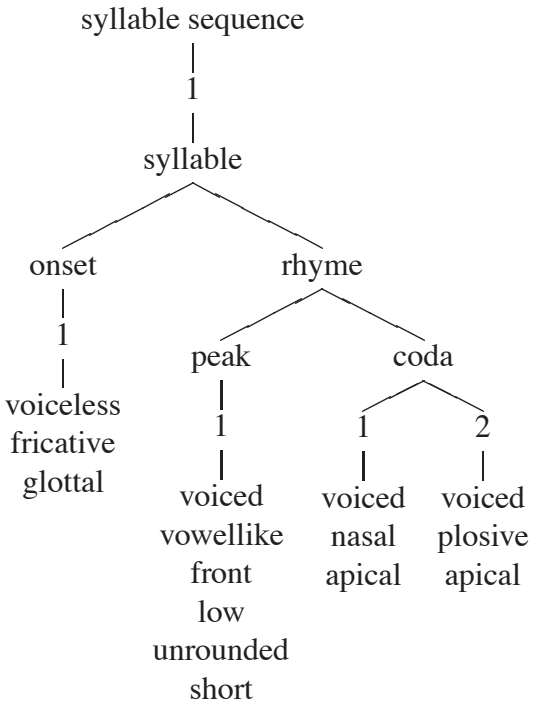

Figure 1: Wordform structure for *hand*

Following (Cahill, 1990), we assume that each syllable consists of an onset (the initial consonant cluster) and a rhyme, and the rhyme consists of a peak (the vowel(s)) and a coda (the final consonant cluster). Each vowel or consonant phoneme is represented by a full feature set: for example, the first phoneme of the onset of *hand* is a voiceless fricative glottal consonant. Where sequences of components (syllables, phonemes etc.) occur, they are numbered from left to right. In the case of *hand*, the syllable sequence contains a single syllable, labelled '1', and the onset and peak of that syllable contain single phonemes, but in the coda we have two phonemes

labelled '1' and '2'. In our model, multiple peaks are only used to describe diphthongs. Long vowels are considered as single peaks.

In figure 1, the analysis consisted of a single root represented as a (1 element) syllable sequence. More complex word forms are described using a binary concatenation node 'concat'. We view such concatenations as themselves syllable sequences (but not flat ones), so that we have a cyclic representation space: concat can dominate flat syllable sequences (primitive units of word formation) or other concat nodes, allowing more complex word forms to be represented. Note however that we do not distinguish prefixation from suffixation, or roots from affixes: concat is simple left-right concatenation. Figure 2 shows the syllable sequence structure for the English plural form *fingers*, consisting of the concatenation of a two element sequence *finger* and the single element sequence *s*[2]. Notice that this last element would not conventionally be considered a complete phonological syllable, but at this level of the analysis it is useful to treat it in the same way as other 'real' syllables. We return to this point in section 6 below.

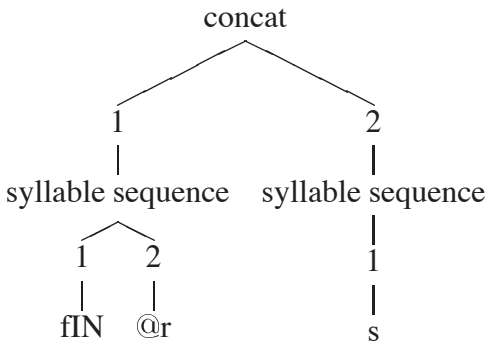

Figure 2: Word form structure for *fingers*

4. The organisation of lexemes

The previous section discussed the structure of the word forms the lexicon aims to represent. In this section we outline the way default inheritance and rule application are used to represent such word forms compactly.

The framework defines word form trees *equationally*, that is the trees do not exist as data structures (attribute-value matrices, for example), but are described by sets of equations each of which associates a path from root to leaf in a tree with one of the phonological feature values specified at that leaf. For example, the 'onset' subtree in figure 1 can be described by the following three equations:

```
<phon 1 syll onset 1 phonation> == voiceless
<phon 1 syll onset 1 manner> == fricative
<phon 1 syll onset 1 place> == glottal
```

These equations are not simply listed for each word form, however, but are organised into an inheritance hierarchy. There are three main components to the organisational structure of the lexicon.

[2]Here and sometimes below we use SAMPA CELEX transcriptions (Baayen et al. 1995, Wells 1987,1989) to abbreviate actual phonological feature bundles where the feature details are not important.

First, the framework is **lexeme-based**: the 'objects' in the lexicon correspond to families of words, and individual word forms are obtained by applying 'morphosyntactic' functions such as `singular, plural, present, past, nominal, gerund`. Such functions can be combined to produce further word forms: for example, applying `nominal+plural` to the lexeme for *Love* gives the structure for the word form *lovers*. The 'base' form (with no functions applied) does not correspond to any word form (although we could choose to make it do so).

Second, the lexeme definitions are represented using **default inheritance**: lexemes are defined in terms of more abstract classes (such as *Noun, Verb, Modal_verb* etc.), inheriting information from them, but also overriding inherited information when required. A typical lexeme needs to specify explicitly its basic own phonological structure, but can inherit all the information that specifies how it forms a plural, or nominal, or genitive etc.. If it happens to have, say, an irregular plural, it can specify this itself too, overriding just that part of the inherited information.

Third, the internal components of a phonological form are organised into their own independent inheritance hierarchies: the lexeme hierarchy inherits actual word forms from a syllable sequence hierarchy, which inherits individual syllable structures from a syllable hierarchy, which inherits individual phonemes from the phoneme hierarchy. Figure 3 illustrates these relationships for the Dutch lexeme *Gebed* (prayer). As a lexeme, *Gebed* is primarily linked into the lexeme hierarchy, inheriting from *Noun_EN*, a subclass of *Noun*. But it inherits part of its content, namely its phonological form, from *GEBED* in the syllable sequence hierarchy. *GEBED* is primarily a *Disyllable*, but it inherits part of its content, namely the two syllables it contains, from *GE* and *BED* in the syllable hierarchy. Finally the syllable *BED* inherits part of its structure, from the consonants *b* and *d* and the vowel *E* in the phoneme hierarchy.

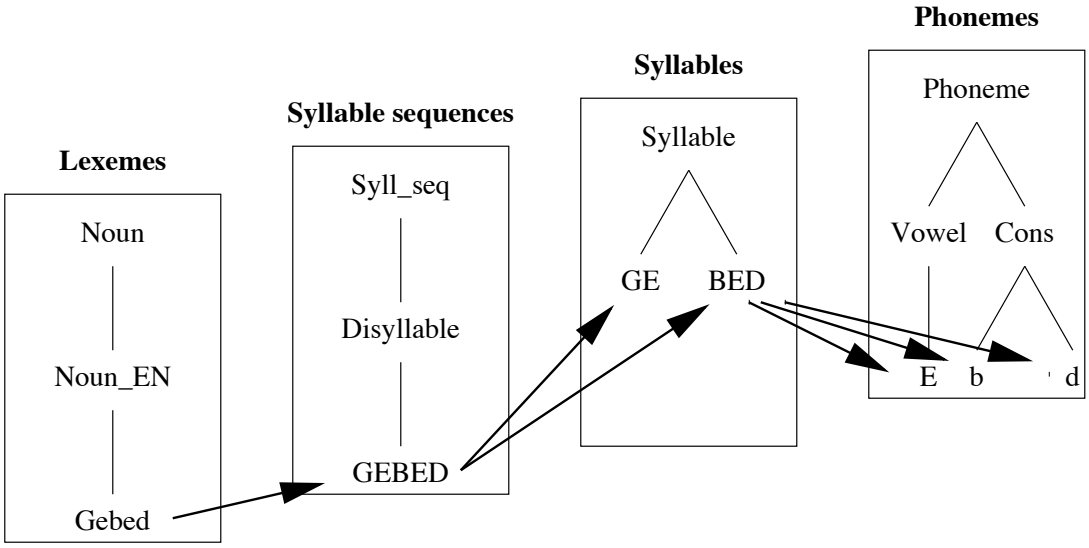

Figure 3: Module and node structure for lexeme Gebed

The lexeme access functions are implemented using the **lexical rule** techniques described in Evans, Gazdar and Weir (1995; 2000), Smets and Evans (1998). Each rule maps an 'input' phonological tree onto an 'output' phonological tree. Each lexeme defines a base tree (which does not correspond to any actual word form – see above), and the first rule applies to that, the second to the output of the first and so on – the final word form is the output of the last rule in

the chain.

Rule definitions are specified in the same inheritance structures as the phonological tree equations. This means that the inheritance hierarchy can be used to control the scope of applicability of a rule: a universal rule can be defined at the top-most node of the hierarchy, one which only applies to nouns at the *Noun* node, one that applies to verbs at the *Verb* node etc. Individual words can even have their own rules if required. In addition, rule definitions can inherit and override from ancestor nodes just as lexeme definitions can: the *Noun_EN* node inherits the *plural* rule from *Noun*, retaining the fact that pluralisation is achieved by concatenating something to the root, but overriding *what* is concatenated - *-en* instead of -s.

Rules may include conditional constructs, testing properties of their input word form structures to decide whether or how to apply. This can be used to control the scope of rule application, for example blocking *superlative* on adjectives that are already in the comparative form, or to control the effect of rule application, for example choosing between /s/ and /z/ as a plural suffix in English. More fundamentally, it is also often used to control *where* in a word form the rule applies. Because word forms are defined equationally, each rule operates on every equation of the word form definition. For most of these it will do nothing, being activated only, for example, in equations relating to the last syllable, or the first peak vowel.

Finally, note that rules can be defined in each of the submodule hierarchies (as shown in figure 3) independently. Rules in the lexeme hierarchy are invoked directly on lexical access, but other rules can be invoked from the hierarchy above them, as we shall see in the examples below.

5. Illustration of the framework

In this section we illustrate these aspects of our framework with some examples from Dutch, English, German, and Danish.

First, we consider an example of multilingual rule definition – the rule for nouns with a plural ending in -s in English, Dutch, and German. In our framework, this fact will be captured by a lexical rule which adds an *-s* suffix to the root of the noun. This *s* suffix is realised as an /s/ in Dutch and German, and as an /s/ or /z/ in English due to voicing alternation. In all three languages a vowel is inserted before the *-s*, if the root ends in a sibilant. This vowel is realised as a /@/ in Dutch and German, and as an /I/ in English. The definition of the *plural_s* rule in our framework is illustrated schematically below.

Dutch and German inherit the plural_s rule as it is defined in the common part. The English plural_s rule inherits from the common part, but overrides the value of the suffix peak, using /@/ instead of /I/. It also adjust the voicing of the final /s/ depending on the voicing of the preceding phoneme.

A second example considers *final devoicing* in Dutch. Final consonant devoicing applies to root final obstruents (plosives and fricatives) when the root is not inflected or when an inflectional suffix is added which does not begin with a vowel. This is achieved by means of a lexical rule which ultimately just sets the `phonation` feature of the last coda to `voiceless`.

The devoicing rule in the lexeme module invokes *devoicelastsyll* in the syllable sequence module. This does nothing except in the last syllable of the word form, in which it invokes *devoicelastcoda* in the syllable module. This also does nothing except in the last coda of the

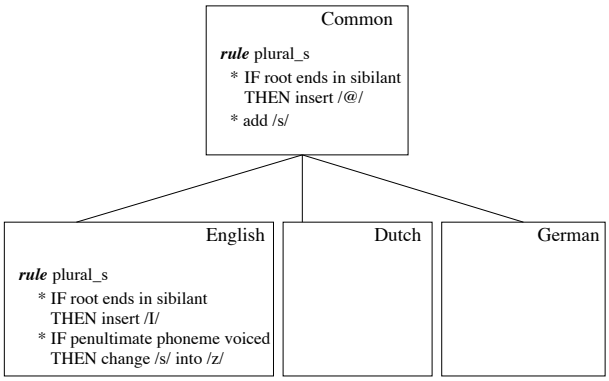

Figure 4: Definition of plural_s rule

syllable, in which it invokes *devoice* in the phoneme module. Here it changes the value for phonation from voiced to voiceless. This process is schematically represented below.

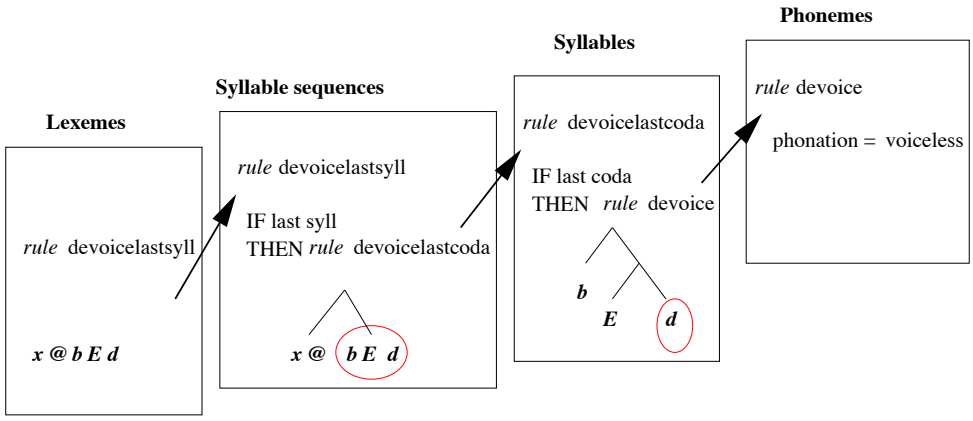

Figure 5: Final devoicing applied to the lexeme Gebed

Notice that rule definitions at each level are not context dependent: *devoicelastcoda* can devoice the last coda of any syllable, not just the last one, and *device* can devoice any phoneme.

In a final example, we look at how a series of rules can be invoked together, using the example of Danish nouns. As in most languages, Danish nouns can occur in singular and plural, but in addition a definite article can be added to the end of the singular or plural form, e.g. *mund* (mouth) - *munden* (the mouth); *munder* (mouths) - *munderne* (the mouths). In the singular the definite article depends on the gender of the noun: *-et* is added to neuter nouns, *-en* to non-neuter nouns. In the plural, first the ending *-er* is added, followed by the definite article *-ne*. The operation of these two rules together in our framework is sketched in figure 6[3] .

6. Conclusions and further work

We have described a framework for lexical representation that is intended to provide a basis for a lexical knowledge base which is both theoretically and computationally satisfactory.

[3] /O_o/ stands for lowered /O/.

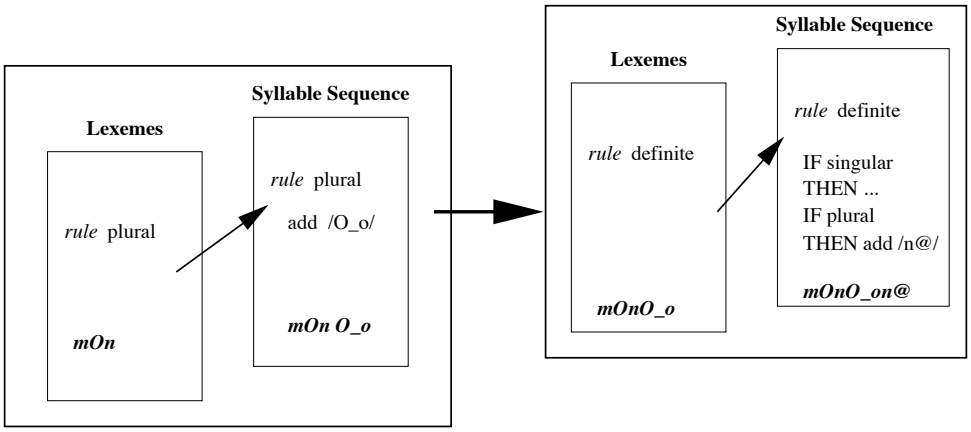

Figure 6: Plural and definite rules applied to the Danish lexeme *Mund* /mOn/ mouth

Although the representation crosses traditional linguistic boundaries, the framework successfully accommodates a wide range of morphological and phonological phenomena in a uniform manner. We make extensive use of inheritance to capture generalisations, motivated entirely by regularities in the word form structures, rather than preconceptions such as 'morpheme' and 'phoneme'. Nevertheless we find that we can identify powerful and linguistically interesting generalisations.

Our principal areas of ongoing development and interest in this research are:

- extending the coverage, both of lexemes and lexical phenomena (phonological rules, inflectional classes etc.). Currently, the framework has been used to describe small sample fragments covering nouns and adjectives in Danish, Dutch, English, and Icelandic. The DATR files of these fragments are available on `http://www.itri.brighton.ac.uk/~Carole.Tiberius/mlex.html`;
- extending the representation, upward to support derivation and compounding, and downward to nonsegmental phonology, phonetics, prosody etc.;
- introducing phonotactics;
- introducing support for orthographic representations;
- interfacing the lexical knowledge base to practical language engineering applications.

A final area of particular interest to us is the issue of *resyllabification rules*. We noted above that our current representation includes 'syllables', such as *s* that are not phonologically valid. Ideally, to construct a 'proper' word form structure these need to be merged into their adjacent syllables. Similarly other syllable boundary phenomena, such as maximum onset, should be addressed. The representation is sufficiently powerful to support such rearrangements of the structure, and so one way to achieve this would be simply to introduce them as obligatory final rules, applied to all word forms.

This approach is appealing because it avoids the need to formally distinguish a level of 'protosyllable' representation for these incomplete structures, making it easier, for example to treat resyllabified word forms as themselves inputs to later word formation processes. However, currently, this remains a topic for future exploration.

References

ANDERSON S. (1988). Morphological theory. In F. NEWMEYER, Ed., *Linguistics: the Cambridge survey*, volume 1, 146–191. Cambridge University Press.

BIRD S. & KLEIN E. (1990). Phonological events. *Journal of Linguistics*, **26**.

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

CAHILL L. (1990). *Syllable-based morphology for natural language processing*. PhD thesis, School of Cognitive Science and Computing Sciences, University of Sussex, Brighton. also available as technical report CSRP 181.

CAHILL L. (1993). Morphophonology in the lexicon. In *Proceedings of the Fifth European Conference on Computational Linguistics*, 87–96.

CAHILL L. & GAZDAR G. (1997). The inflectional phonology of german adjectives, determiners and pronouns. *Linguistics*, **35**(2), 211–245.

CAHILL L. & GAZDAR G. (1999a). German noun inflection. *Journal of Linguistics*, **35**, 1–42.

CAHILL L. & GAZDAR G. (1999b). The PolyLex architecture: multilingual lexicons for related languages. *Traitement Automatique des Langues*, **40**(2), 5–23.

CHOMSKY N. & HALLE M. (1968). *The sound pattern of English*. New York: Harper and Row.

COLEMAN J. (1992). Synthesis by rule without segments or rewrite rules. In C. BENOIT & G. BAILLY, Eds., *Talking Machines*. Elsevier.

CORBETT G. & FRASER N. (1993). Network morphology: A DATR account of russian nominal inflection. *Journal of Linguistics*, **29**, 113–142.

EVANS R. & GAZDAR G. (1996). DATR: A language for lexical knowledge representation. *Computational Linguistics*, **22**(2), 167–216.

EVANS R., GAZDAR G. & WEIR D. (1995). Encoding lexicalized tree adjoining grammars with a nonmonotonic inheritance hierarchy. In *Proceedings of ACL95*, 77–83.

EVANS R., GAZDAR G. & WEIR D. (2000). 'Lexical rules' are just lexical rules. In A. ABEILLE & O. RAMBOW, Eds., *Tree Adjoining Grammars: linguistic, formal and computational properties*, CSLI Lecture Notes. University of Chicago Press.

GIBBON D. (1992). ILEX: a linguistic approach to computational lexica. In U. KLENK, Ed., *Computatio Linguae: Aufsatze zur algorithmischen und quantitativen Analyse der Sprache*, volume Beiheft 73 of *Zeitschrift für Dialektologie und Linguistik*, 32–53. Stuttgart: Franz Steiner.

MOHANAN K. (1986). *Lexical Phonology*. Dordrecht: D. Reidel.

SMETS M. & EVANS R. (1998). A compact encoding of a DTG grammar. In *Proceedings of the 4th Workshop on Tree Adjoining Grammars and Related Formalisms*, University of Pennsylvania, Philadelphia, 164–167.

STUMP G. (1993a). On rules of referral. *Language*, (69), 449–479.

STUMP G. (1993b). Position classes and morphological theory. In G. BOOIJ & J. VAN MARLE, Eds., *Year Book of Morphology 1992*, 129–180. Dordrecht: Kluwer.

ZWICKY A. (1985). How to describe inflection. *BLS*, (11), 371–386.

ZWICKY A. (1990). Inflectional morphology as a (sub)component of grammar. In W. U. DRESSLER, H. C. LUSCHÜTZKY, O.E. PFEIFFER, & J.R. RENNISON, Ed., *Contemporary Morphology*, 217–236. Berlin: Mouton de Gruyter.