Incorporating Metaphonemes in a Multilingual Lexicon

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

Abstract

This paper describes a framework for multilingual inheritance-based lexical representation which allows sharing of information across languages at all levels of linguistic description. The paper focuses on phonology. It explores the possibility of establishing a phoneme inventory for a group of languages in which language-specific phonemes function as "allophones" of newly defined *metaphonemes*. Dutch, English, and German were taken as a test bed and their vowel phoneme inventories were studied. The results of the cross-linguistic analysis are presented in this paper. The paper concludes by showing how these metaphonemes can be incorporated in a multilingual lexicon.

1 Introduction

This paper describes a framework for multilingual inheritance-based lexical representation which allows sharing of information across (related) languages at all levels of linguistic description. Most work on multilingual lexicons up to now has assumed monolingual lexicons linked only at the level of semantics (MULTILEX 1993; Copestake et al. 1992). Cahill and Gazdar (1999) show that this approach might be appropriate for unrelated languages, as for example English and Japanese, but that it makes it impossible to capture useful generalisations about related languages – such as English and German. Related languages share many linguistic characteristics at all levels of description – syntax, morphology, phonology, etc. – not just semantics. For instance, words which come from a single root have very similar orthographic and phonological forms. Compare English, Dutch, and German[1]:

English	**Dutch**	**German**
bed	*bed*	*Bett*
/bEd/	/bEt/	/bEt/
rib	*rib*	*Rippe*
/rIb/	/rIp/	/rIp@/
hand	*hand*	*Hand*
/h{nd/	/hAnt/	/hant/
cat	*kat*	*Katze*
/k{t/	/kAt/	/kats@/

Most differences can be attributed to different orthographic conventions and regular phonological changes (e.g. final devoicing in Dutch and German). The English /{/, the Dutch /A/, and the German /a/ in the last two examples, are even virtually the same. They have slightly different realisations but they are phonologically non-distinctive, i.e. if the Dutch /A/ were substituted by the English /{/ in Dutch, the result would not be a different word, but it would simply sound like a different accent.

Cahill and Gazdar (1999) describe an architecture for multilingual lexicons which aims to encode and exploit lexical similarities between closely related languages. This architecture has been successfully applied in the PolyLex project[2] to define a trilingual lexicon for Dutch, English, and German sharing morphological, phonological, and morphophonological information between these languages.

In this paper, we will take the PolyLex framework as our basis. We will focus on the phonological similarities between related languages and we will extend the PolyLex approach by capturing cross-linguistic phoneme correspondences, such as the /{/ - /A/ - /a/ correspondence mentioned above[3].

First, we will discuss how a phoneme inventory can be defined for a group of languages – Dutch,

[1]The transcriptions are taken from CELEX (Baayen et al. 1995) and use the SAMPA phonetic alphabet (Wells 1989).

[2]http://www.cogs.susx.ac.uk/lab/nlp/polylex/

[3]We believe the approach would be even more beneficial if extended to a featural level, but for the present purposes we confine ourselves to the segmental level.

English, and German. Then, we will explain the multilingual architecture used in PolyLex. Finally, we will explore how these cross-linguistic phoneme correspondences can be integrated into the multilingual framework.

2 A Metaphoneme Inventory

In this section we describe how a phoneme inventory can be defined for a group of languages in which language-specific phonemes function as "allophones" of newly defined metaphonemes. We will restrict ourselves to the vowel phonemes of Dutch, English, and German. If we know, for example, that words which are realised with an /{/ in English are usually realised with an /A/ in Dutch, and an /a/ in German (as in *hand* /h{nd/ versus /hAnt/ versus /hant/, *cat* /k{t/ versus /kAt/ versus /kats@/, etc.), we might be able to generalise over these three language-specific phonemes and introduce a metaphoneme, e.g. |{Aa|, which captures this generalisation.

To give an impression of the distribution of the different vowel phonemes across Dutch, English, and German, their vowel charts (König and van der Auwera 1994; Wells 1989) were merged into one big vowel chart containing all the vowel phonemes of these three languages.[4], The resulting chart is given in figure 1[5]:

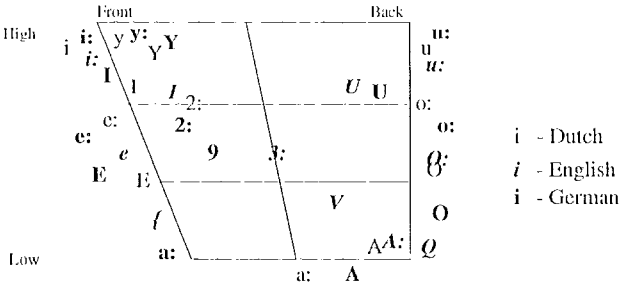

Figure 1: Vowel phonemes in Dutch, English, and German

This figure shows which vowel phonemes are realised in which language (e.g. /{/ occurs in English, but not in Dutch and German), but it does not tell us anything about cross-linguistic phoneme correspondences. Knowing that Dutch and German both have a phoneme /o:/, does not mean that they are cross-linguistically non-distinctive.

To find cross-linguistic phoneme correspondences, we followed O'Connor's (1973) strategy for establishing phoneme correspondences between different accents, identifying phonemes of one accent with those of another:

> "How are we to decide whether to equate phoneme X with phoneme A or with phoneme D? We can do so only on the basis of the words in which they occur: if X and A both occur in a large number of words common to both accents we link them together as representing the same point on the pattern. If, on the other hand, X shares more words with D than with A, we link X and D. [...] Even so, if X and D occur in a very similar word-set and X and A do not, then it is much more revealing to equate X and D than X and A."
> (O'Connor 1973, p.186)

We extended O'Connor's strategy and applied it to a group of (closely) related languages sharing a common word stock – in our case a subset of the West Germanic languages sharing words with a common Germanic origin. We compiled a list of 800 (mono- and disyllabic) Germanic cognates, looked up the transcriptions in the CELEX database (Baayen et al. 1995), and then mapped words containing a particular vowel in one language onto its cognates in the other two languages to see how this particular vowel was realised in the other two languages. This process was repeated for all the vowels, for all three languages.

A few examples of the results we obtained for English vowels are included below[6].

As can be seen from these[7], there is some variation in the closeness of the correspondences. The vowel set /{/ - /A/ - /a/, as we anticipated at the outset, does turn out to be a valid correspondence. The set associated with English /i:/, on the other hand, is less clearcut, as there are several possible cor-

[4]Phonemes that only occur in loanwords were not included as languages adapt loanwords to different degrees to their own phonetic system.

[5]The vowels are described along the three dimensions of vowel quality: [high], [back], and [round]. The rounded vowels are /y,**y**:,Y,**Y**,2:,**2:**,9,*Q*,**O**,O, *O:*,**o:**,o:,u,**u:**,*u:*,*U*,**U**/.

[6]The remaining correspondence tables are available at `http://www.itri.bton.ac.uk/~Carole.Tiberius/mphon.html`

[7]Note that the total number of words is not always exactly the same in all three languages. This is because for some words the corresponding phonemic transcription was not found.

English		Dutch		German	
{	37	A	27	a	22
		a:	3	a:	3
		E	2	E	3
		}	2	I	2
		o:	2	e:	1
		u:	1	O	1
				o:	1
				u:	1
				|:	1
		total	37	total	35

Table 1: Correspondences for English /{/ words as in *hand* /h{nd/ vs /hAnt/ vs /hant/.

English		Dutch		German	
i:	65	a:	14	a:	12
		o:	11	i:	8
		e:	9	ai	7
		i:	8	e:	5
		u:	7	y:	5
		I	5	au	5
		E	4	I	5
		EI	3	o:	4
		|:	2	a	3
		/I	1	E	3
		A	1	u:	3
				O	2
				E:	1
				Y	1
				|:	1
		total	65	total	65

Table 2: Correspondences for English /i:/ words as in *meal* /mi:l/ vs /ma:l/ vs /ma:l/ and *deep* /di:p/ vs /di:p/ vs /ti:f/.

responding vowel phonemes in the other two languages. If we consider the correspondences from the starting point of one of the other languages, the results are slightly different. For instance, English /A:/ corresponds strongly to Dutch /A/, but Dutch /A/ corresponds almost equally to English /{/ and /A:/. Further investigation is required to ascertain how many of these cases can be further generalised by recourse to phonological or phonotactic properties of the words in question. Currently the mapping from metaphoneme to (language-specific) phoneme requires reference only to the language. For a more

English		Dutch		German	
A:	31	A	19	a	15
		a:	4	a:	5
		E	4	E	5
		O	2	e:	2
		e:	1	E:	1
		EI	1	U	1
				Y	1
				ai	1
		total	31	total	31

Table 3: Correspondences for English /A:/ words as in *heart* /hA:T/ vs /hArt/ vs /hart/.

Dutch		English		German	
A	77	{	25	a	53
		A:	17	a:	9
		eI	10	E	6
		O:	8	I	3
		Q	4	ai	1
		@U	4	e:	1
		u:	2		
		E	2		
		3:	2		
		i:	1		
		I	1		
		aI	1		
		total	77	total	73

Table 4: Correspondences for Dutch /A/ words as in *hand* (hand) and *hart* (heart).

sophisticated analysis, phonological and phonotactic information would need to be considered as well. However, even at the present level of analysis, the metaphoneme principle can be helpful in the multilingual lexical structure proposed, as we now discuss.

3 The multilingual inheritance lexicon

In this section, we will explore the sharing of phonological information in the lexical entries of a multilingual inheritance-based lexicon. We focus on phonology rather than orthography as phonology is nearer to primary language use (i.e. spoken language), it can be used as input for hyphenation rules, spelling correction, and it is essential as the level of symbolic representation for speech synthesis (MULTILEX 1993).

We will take the multilingual architecture of PolyLex as our starting point. First, we will describe the PolyLex architecture. Then, we will show how phonological information can be shared in the lexical entries.

PolyLex defines a multilingual inheritance-based lexicon for Dutch, English and German. It is implemented in DATR, an inheritance-based lexical knowledge representation formalism (Evans and Gazdar 1996). The rationale of inheritance-based lexicons requires information to be pushed as far up the hierarchy as it can go, generalising as much as possible. In a multilingual lexicon, this means that information which is common to several languages is stated at higher points in the hierarchy than that which is unique to just one of the languages. In addition, PolyLex makes use of orthogonal multiple inheritance which allows a node in the hierarchy to inherit different kinds of information (e.g. semantics, morphology, phonology, syntax) from different parent nodes. In this paper, we are just interested in the phonological hierarchy.

PolyLex assumes a contemporary phonological framework in which all lexical entries are defined as having a phonological structure consisting of a sequence of structured syllables, a syllable consisting of an onset (the initial consonant cluster, which might be split up into onset 1, onset 2, etc.) and a rhyme. The rhyme consists of a peak (the vowel) and a coda (the final consonant cluster, which might be split up into coda 1, coda 2, etc.). This structure is defined at the top of the hierarchy, and applies by default to all words. Only the relevant values for onset, peak, and coda have to be defined at the individual lexical entries (see Cahill and Gazdar 1997). Following PolyLex we will concentrate on a segmental phonemic representation. An example of the lexical entry *gram* as it would be represented in PolyLex, is shown in figure 2.

The multilingual phonological entry for *gram* is defined by sharing identical segments occurring in the majority of the language-specific entries (/gr{m/ – /xrAm/ – /gram/). That is, *onset 1* is /g/, *onset 2* is /r/, and *coda* is /m/.

English and German can inherit all the information from the common part except for the value of their peak, which is respectively /{/ and /a/. In Dutch, the value of the peak has to be specified as being /A/, plus we will have to override the value for the first onset to get [xrAm].

This example misses the generalisation that the

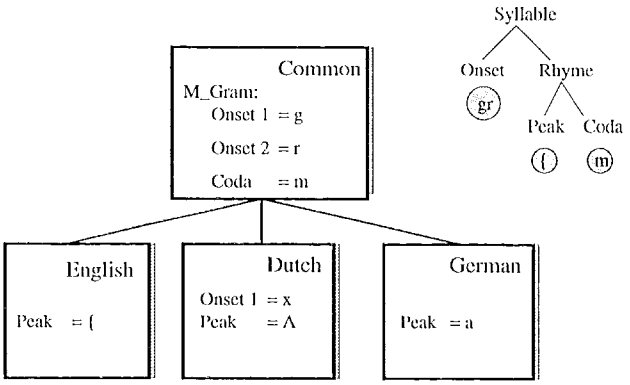

Figure 2: A multilingual inheritance lexicon without metaphonemes

English /{/, the Dutch /A/, and the German /a/ are phonologically non-distinctive. For each lexical entry where English uses /{/, Dutch /A/, and German /a/, the value for peak has to be specified in the language-specific parts. By using the metaphoneme |{Aa| instead, this information needs to be specified only once. The resulting multilingual phonemic representation for *gram* is given in figure 3.

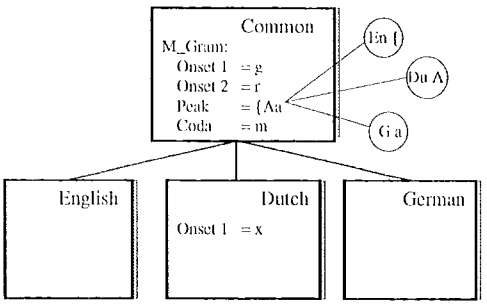

Figure 3: A multilingual inheritance lexicon with metaphonemes

All the information has now been pushed up as far as it can go, capturing as many generalisations as possible. The information that |{Aa| results in an /{/ in English, an /A/ in Dutch, and an /a/ in German is specified only at the top level. The language-specific boxes are almost empty, except for the value of the first onset in Dutch. The reason for this is that as yet we have only defined cross-linguistic phoneme correspondences for vowels, not for consonants. We do, however, suspect that the Dutch /x/ is phonologically non-distinctive from the German and English /g/. Further research defining cross-linguistic phoneme correspondences for consonants

will have to confirm this.

It is a fundamental feature of this account that the inherited information is only *default* information which can be overridden. Thus, it is not required that metaphoneme correspondences are complete and we may choose to use a metaphoneme even if one of the languages uses a different vowel in some words. The definitions can be overridden in exactly the same way as the onset definition in Dutch in the example above. So if we consider the vowel correspondences in table 1, we can see that of the 35 words which have cognates in all three languages, 27 can be defined as having the metaphoneme |{Aa| in the common lexical entry (those for which both English and Dutch have the corresponding vowels). Five of these will require a separate vowel defined for German, while the remainder will need separate vowel definitions for all three languages.

Given this, we can see that economy of representation can be achieved even in cases where the vowel correspondences are far from conclusive. Even if only half or fewer of the Dutch words, for example, have the same vowel in cognates for which the English words have the same vowel, this still means that those half can be defined without the need for the language-specific vowel to be defined.

Another feature of the metaphoneme principle that differentiates it from the phonemic principle is that there is no requirement for biuniqueness. A phoneme in a language can be a realisation of more than one metaphoneme. This means that we can define a metaphoneme |{Aa| as well as another, |A:Aa|. Each of these will then be used in different common lexical entries. This can be used as an alternative to phonological/phonotactic conditioning or in addition to it, for just those cases where there is more than one correspondence but no obvious phonological/phonotactic conditioning for the decision between phonemes.

4 Conclusion

In this paper, we have discussed the concept of *metaphonemes*. Metaphonemes are cross-linguistic phoneme correspondences such as the English /{/, the Dutch /A/, and the German /a/ correspondence mentioned above. At the multilingual level, the realisation of the metaphoneme is conditioned by the choice of language. At the lower monolingual level its realisation as an allophone of a particular phoneme is conditioned by the phonological environment. As such, a metaphoneme is a generalisation of a generalisation.

We have shown how a metaphoneme inventory can be defined for a group of languages and that incorporating these cross-linguistic phoneme correspondences in a multilingual inheritance lexicon increases the number of generalisations that can be captured. Calculations on the syllable inventories of Dutch, English, and German in the CELEX database show that the introduction of metaphonemes increases the amount of sharing at the syllable level by about 25%.

Another benefit of introducing metaphonemes is improved robustness in NLP systems. Knowledge about cross-linguistic commonalities can help to provide grounds for making an "intelligent" guess when a lexical item for a particular language is not present.

This research has concentrated on cross-linguistic vowel phoneme correspondences. Similar research will be done for consonants.

References

Baayen, H., R. Piepenbrock and H. van Rijn. 1995. *The CELEX Lexical Database*, Release 2 (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Cahill, L. and G. Gazdar. 1997. "The inflectional phonology of German adjectives, determiners and pronouns", In *Linguistics*, 35.2, pp.211-245.

Cahill, L. and G. Gazdar. 1999. "The PolyLex architecture: multilingual lexicons for related languages", In *Traitement Automatique des Langues*, 40:2, pp.5-23.

Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni, and E. Marinai. 1992. "Multilingual Lexical Representation". *ESPRIT BRA-3030 ACQUILEX Working Paper* N° 043.

Evans, R. and G. Gazdar. 1996. "DATR: A Language for Lexical Knowledge Representation", In *Computational Linguistics*, Vol. 22-2, pp.167-216.

König, E. and J. van der Auwera (eds.) 1994. *The Germanic Languages*, Routledge, London.

MULTILEX, 1993. "MLEX$_d$ Standards for a Multifunctional Lexicon", Final Report, CAP GEMINI INNOVATION for the MULTILEX Consortium, Paris.

O'Connor, J.D. 1973. *Phonetics*, Pelican Books, Great Britain.

Wells, J. 1989. "Computer-coded phonemic notation of individual languages of the European Community", In *Journal of the International Phonetic Association*, 19:1, pp.31-54.