A MetaPhoneme Inventory

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 focuses on the sharing of phonological information in a multilingual inheritance-based lexicon. 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*. Danish, 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 inheritance-based lexicon.

1 Introduction

The work described here assumes 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 (1995;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.

Capturing such similarities can help to produce more robust, more readily maintainable and more readily extensible multilingual natural language processing systems for related languages (Cahill and Gazdar 1995;1999). Consider lexical incompleteness. The multilingual inheritance architecture with cross-linguistic information sharing allows one to exploit default information from both source and target languages together with information about the default commonalities across those languages. This way it may be possible to deduce sufficient information about a missing lexical item via information which is available in the lexicon. Imagine that we want to know the German word for *forbid*, but this word is not in our lexicon. Assume, however, that the lexicon contains the English verb *bid* and its German equivalent *bieten*. In addition, our lexicon may

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

know that verbs beginning with the syllable *for* in English generally start with *ver* in German. The English verb *forgive*, for example, has the German equivalent *vergeben*, the English verb *forget* has the German equivalent *vergessen*, etc. On the basis of this information, it is possible to construct a hypothesised German form by simply adding the syllable *ver* onto the verb *bieten*, giving the form *verbieten*. In this case, the hypothesised form is the correct translation of *forbid*. This will not always be the case because of lexical idiosyncrasies to be found in one or both languages. This kind of educated guess is, however, the best we can do given the way English and German work and given the way they usually relate to each other (Cahill and Gazdar 1995).

Cahill and Gazdar (1995;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 to define a trilingual lexicon for Dutch, English, and German sharing morphological, phonological, and morphophonological information between these languages.[2]

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 – Danish, Dutch, English, and German. Then, we will explain the multilingual architecture used in PolyLex. Finally, we will discuss the advantages of integrating these cross-linguistic phoneme correspondences 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 four Germanic languages – Danish, 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 and Danish (as in *cat* /k{t/ versus /kAt/ versus /kats@/ versus /kad/), we might be able to generalise over these four 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 Danish, Dutch, English, and German, their vowel charts (Basböll and Wagner 1985; König and van der Auwera 1994; Wells 1989;1995) were merged into one big vowel chart containing all the vowel phonemes of these four languages.[4] The resulting chart is given in figure 1:[5]

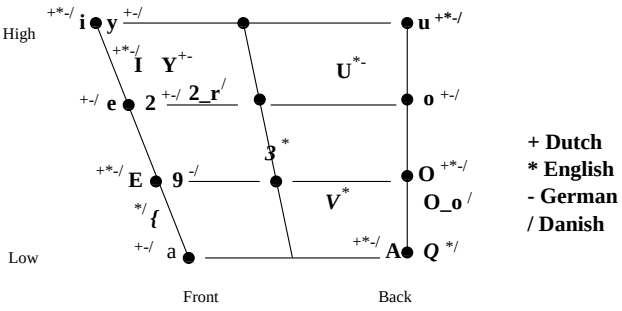

Figure 1: Vowel phonemes in Danish, Dutch, English, and German

This figure shows which vowel phonemes are realised in which language (e.g. /{/ occurs in English and Danish, 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

[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.

[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,2,2_r,9,Q,O,o,U,u/. All Danish vowels in this chart can be either long or short.The extension "_r" means that the vowel is raised, "_o" means that the vowel is lowered.

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)

For example, O'Connor (p.187) compares Yorkshire and British English Received Pronunciation (RP), and concludes that both have the phonemes /E,{,Q/ in opposition in largely the same set of words, *pet*, *pat*, *pot*, and that in addition there is a set of words all of which have /U/ in Yorkshire, but some of which have /V/ and some /U/ in RP. For instance, both *but* /bVt/ and *put* /pUt/ in RP will be realised with the /U/ phoneme in Yorkshire resulting in respectively /bUt/ and /pUt/ . Thus, Yorkshire /U/ can be linked to both RP phonemes /V/ and /U/. We capture this situation by introducing a metaphoneme |UV| for those words which have /U/ in Yorkshire but /V/ in RP, in addition to the phonemes /E,{,Q,U/ which occur in both accents in largely the same set of words.

For our research purposes, 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 Germanic languages sharing words with a common Germanic origin. We compiled a list of 800 (mono- and disyllabic) Germanic cognates, looked up the transcriptions (Baayen et al. 1995, Hansen 1990), and then mapped words containing a particular vowel in one language onto its cognates in the other three languages to see how this particular vowel was realised in the other three languages. This process was repeated for all the vowels, for all four languages.

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

English		**Dutch**		**German**		**Danish**	
{	37	A	27	a	22	a	8
		a:	3	a:	3	A	6
		E	2	E	3	a:	3
		}	2	I	2	e	3
		o:	2	e:	1	O:	1
		u:	1	O	1	O_o	1
				o:	1	y:	1
				u:	1	Q:	1
				|:	1	o:	1
						2	1
		total	37	total	35	total	26

Table 1: Correspondences for English /{/ words as in *cat* /k{t/ vs /kAt/ vs /kats@/ vs /kad/.

English		**Dutch**		**German**		**Danish**	
i:	65	a:	14	a:	12	E:	4
		o:	11	i:	8	2_r:	3
		e:	9	ai	7	2_r	3
		i:	8	e:	5	y:	3
		u:	7	y:	5	E	2
		I	5	au	5	A	2
		E	4	I	5	O:	2
		EI	3	o:	4	9	2
		|:	2	a	3	u:	2
		/I	1	E	3	a:	2
		A	1	u:	3	2	1
				O	2	A:	1
				E:	1	O	1
				Y	1	Q:	1
				|:	1	Q	1
		total	65	total	65	total	30

Table 2: Correspondences for English /i:/ words as in *seed* /si:d/ vs /za:t/ vs /za:t/ vs /sED/ and *deep* /di:p/ vs /di:p/ vs /ti:f/ vs /dy:b/.

As can be seen from these, there is some variation in the closeness of the correspondences depending on language and vowel phoneme.[7]

[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 four languages. This is because for

English		Dutch		German		Danish	
A:	31	A	19	a	15	a	8
		a:	4	a:	5	a:	3
		E	4	E	5	{	3
		O	2	e:	2	A	2
		e:	1	E:	1	i:	1
		EI	1	U	1	O	1
				Y	1	e:	1
				ai	1		
		total	31	total	31	total	19

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

Dutch		English		German		Danish	
A	77	{	25	a	53	a	23
		A:	17	a:	9	A	11
		eI	10	E	6	a:	7
		O:	8	I	3	E	5
		Q	4	ai	1	O_o	3
		@U	4	e:	1	{	3
		u:	2			A:	2
		E	2			Q:	2
		3:	2			e	2
		i:	1			E:	1
		I	1			O:	1
		aI	1			i:	1
						o:	1
		total	77	total	73	total	62

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

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 corresponding vowel phonemes in the other three languages. Especially in Danish, there is no clear favourite. All vowels in the Danish list have about the same likelihood of occurrence. Overall, the correspondences seem to be less clearcut for Danish than for the other three languages. This is as expected, as Danish is the most distant of the four languages, belonging to the North Germanic language family, while Dutch, English, and German are all West Germanic 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 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. For clarity, we will ignore all other aspects of the lexicon such as semantics, syntax, and morphology, and focus purely on phonology. 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

some words the corresponding phonemic transcription was not found.

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) and a rhyme. The rhyme consists of a peak (the vowel) and a coda (the final consonant cluster). 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 *hair* as it would be represented in PolyLex, is shown in figure 2.

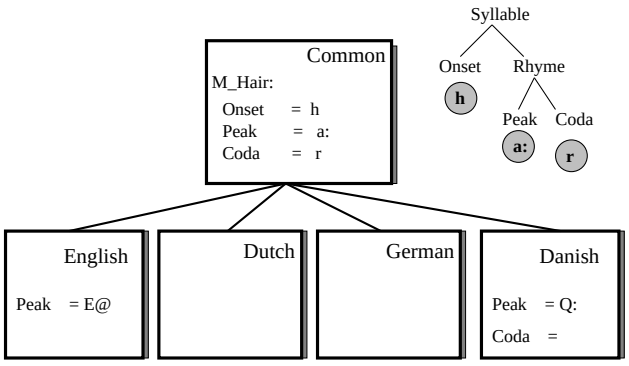

Figure 2: A multilingual inheritance lexicon without metaphonemes

The multilingual phonological entry for *hair* is defined by sharing identical segments occurring in the majority of the language-specific entries (/hE@r/ in English, /ha:r/ in Dutch and German, /hQ:/ in Danish). That is, *onset* is /h/, *peak* is /a:/, and *coda* is /r/.[8]

Dutch and German can inherit all the information from the common part. English and Danish need to override the value of the peak which is respectively /E@/ and /Q:/. In addition, Danish needs to specify that the value of the coda is null.

This example misses the generalisation that the English /E@/, the Dutch and German /a:/, and Danish /Q:/ are phonologically non-distinctive. For each lexical entry where English uses /E@/, Dutch and German /a:/, and Danish /Q:/, the value for peak has to be specified in the language-specific parts. By using the metaphoneme |E@a:Q:| instead, this information needs to be specified only once. The resulting multilingual phonemic representation for *hair* is given in figure 3.

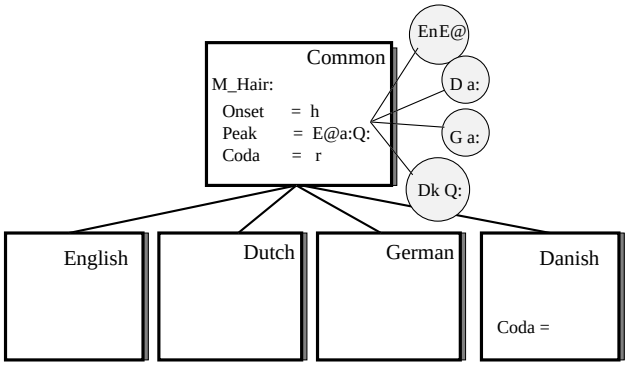

Figure 3: A multilingual inheritance lexicon with metaphonemes

All the information has now been pushed up as far as it can go, capturing as many generalisations as possible. The information that |E@a:Q:| results in an /E@/ in English, an /a:/ in Dutch and German, and an /Q:/ in Danish is specified only at the top level. The language-specific boxes are almost empty, except for the value of the coda in Danish, which is defined as null.

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. So if we consider the vowel correspondences in table 1, we can see that of the 37 words which have cognates in some of the four languages, 27 can be defined as having the metaphoneme |{Aa| in the common

[8] In Standard British English pronunciation, the final /r/ is not always realised. CELEX, however, includes it, and it could be reasonable viewed as an underlying segment.

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 nineteen will require a separate vowel defined for Danish. The remainder of the words will need separate vowel definitions for all four languages. For instance, the lexical entry for *hand* requires a separate vowel for Danish, as can be seen in figure 4 below. As yet we have only defined cross-linguistic phoneme correspondences for vowels, not for consonants. However, the English /d/ and the Dutch and German /t/ are phonologically non-distinctive in syllable final position and this could be captured by introducing a rule which devoices syllable final obstruents in Dutch and German but not in English.

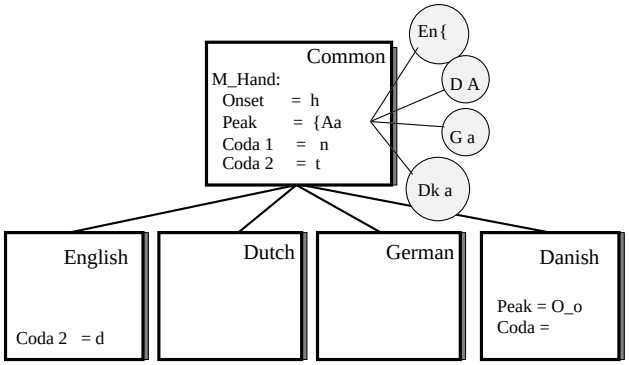

Figure 4: A multilingual inheritance lexicon with metaphonemes

Given the default nature of this information, 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 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.

To support our claims, we compared the syllable inventories for Dutch, English, and German in the CELEX database (the database does not contain data for Danish) and calculated how many syllables they have in common by taking the sum of the overlap of syllables between languages divided by the total number of syllables per language, and then dividing this by the number of languages, i.e.

$$\frac{\sum(overlap\ between\ languages/total\ per\ language)}{number\ of\ languages}$$

The first part of this expression, *overlap between languages/ total per language*, gives the amount of sharing for a single language. The rest of the sum just averages across the number of languages involved.

Let us now calculate the amount of sharing between Dutch, English, and German on the basis of the CELEX database. The CELEX database contains 5193 different syllables for Dutch, 8713 for German, and 7096 for English. 857 of those are shared between the three languages. Applying our formula, this results in

$$\frac{\frac{857}{5193} + \frac{857}{8713} + \frac{857}{7096}}{3} = 0.13$$

This means that 13% of the syllables of Dutch, English, and German in the CELEX database are shared between the three languages. We then did the same calculation but incorporated metaphonemes in the syllable inventories given by CELEX. The amount of sharing rose to 20%. Finally, we calculated the amount of sharing after replacing all vowel phonemes in the syllable inventories by one single vowel phoneme, resulting in 30% sharing.

The latter case is equivalent to the maximal amount of sharing that can be obtained by including metaphonemes, i.e. all vowels correspond to one single metaphoneme. Thus, the inclusion of metaphonemes results in an improvement of 7 out of 30 points, i.e. metaphonemes increase the amount of sharing between Dutch, English, and German at the syllable level by 23.33%.

The potential uses for an approach such as that described here are many and varied. In addition to the possibility the general framework offers for increased robustness in multilingual NL systems (as suggested by Cahill and Gazdar (1995)), the extension of the model to the metaphoneme level can also offer a range of applications in NL and speech systems. As suggested in section 2 above, the approach we suggest for different but closely related *languages* is also applicable to different accents within a single language. Just as we suggest above that a speaker using the wrong phonemic variant of a metaphoneme would sound as though they have a different accent, so the principle could be employed explicitly to produce speech with different accents. Although the work described above is very far from such applications at this stage, there exists the potential to "tune" speech synthesisers to particular languages in a linguistically principled and robust way.

Indeed, our approach to modelling language or dialect *similarity* mirrors the work of Nerbonne et al. (e.g. Nerbonne and Heeringa (1997), Nerbonne et al. (1996)), modelling dialect *dissimilarity*. Their work could be viewed as taking the phonological correspondences that we model, measuring the distance between the realisations of the metaphonemes in order to determine the distance between different dialects.

Another potential area of application for such an approach is in the field of language learning. It is clear that the kinds of substitution errors (where one sound is – usually consistently – replaced by another similar one) that are actually found do not necessarily correspond to metaphoneme correspondences. For example, Dutch speakers, who often have difficulty reproducing the English /{/ segment, tend to replace it with a sound closer to /E/ than to the /A/ that corresponds to it in our metaphoneme inventory. However, it is likely that at least some of the correspondences we propose would be helpful in suggesting the types of errors learners are likely to make and in demonstrating to them the correspondences and distinctions between the phoneme inventories of the different languages.

Within computational linguistics it is possible that the metaphoneme correspondences we suggest could assist in phonology-orthography mapping. In languages like English, where the spelling is based largely on a historical representation of the phonology, it is possible that an underlying representation of phonology that had some historical foundations might be more helpful in determining the orthography. Metaphoneme definitions that distinguish different uses of (synchronically) the same segment might permit easier orthographic correspondences. For instance, in French, the é and è characters are non-distinctive synchronically, but their orthographic distinctions are representative of historical phonological differences which may be represented in the metaphoneme correspondences for French and its closest relatives.

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.

Basböll H. and J. Wagner. 1985. *Kontrastive Phonologie des Deutschen und Dänischen*, Niemeyer Verlag, Tübingen.

Cahill, L. and G. Gazdar. 1995. "Multilingual Lexicons for Related Languages", In *Proceedings of the 2nd DTI Language Engineering Conference*, pp. 169-176.

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^o 043.

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

Hansen, P.M. 1990. *Udtaleordbog*, Gyldendal, Copenhagen.

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.

Nerbonne, J., W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten, and W. van de Vis. 1996. "Phonetic Distance between Dutch Dialects", In G. Durieux, W. Daelemans, and S. Gillis (eds.) *Proceedings of CLIN'95*, Antwerp, pp. 185-202.

Nerbonne, J. and W. Heeringa. 1997. "Measuring Dialect Distance Phonetically", In John Coleman (ed.) *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pp.11-18.

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.

Wells, J. 1995. *Computer-coding the IPA: a proposed extension of* SAMPA, Available anonymous ftp:pitch.phon.ucl.ac.uk in directory /pub/sam/ipasam-x.ps.