Cross-linguistic phoneme correspondences

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

Surrey Morphology Group
University of Surrey
Guildford
UK
c.tiberius@surrey.ac.uk

Abstract

Cross-linguistic phoneme correspondences, or *metaphonemes*[1], can be defined across languages which are relatively closely related in exactly the same way as correspondences can be defined for dialects, or accents, of a single language (e.g. O'Connor, 1973; Fitt, 2001). In this paper we present the theory of metaphonemes, comparing them with traditional archi- and morphophonemes as well as with similar work using "keysymbols" done for accents of English. We describe the metaphoneme inventory defined for Dutch, English and German, comparing the results for vowels and consonants. We also describe some of the unexpected information that arose from the analysis of cognate forms we undertook to find the metaphoneme correspondences.

1 Introduction

Tiberius and Cahill (2000) presented the theory of cross-linguistic phoneme correspondences (metaphonemes) with an example pilot study of the vowels of Dutch, English and German. The aim of this work is to allow the type of generalisation that is permitted by the use of phonemes with allophonic variation to be taken one level higher, i.e. above the level of the single language. The idea behind it is to represent the near-identities that closely related languages such as Dutch, English and German so often share. For example, the Dutch word 'kat' /kAt/ has the English equivalent 'cat' /k{t/, and the German 'Katze' /kats@/[2]. While the consonants are largely identical (/k/–/k/–/k/ and /t/–/t/–/ts/), the vowels are subtly different. However, they are not *distinctive* – i.e. if the /{/ in English were replaced with /a/ it would not sound like a different word, but rather it would sound like a different accent. Thus our aim is not to construct a universal phoneme set representing all phonemes occurring in a particular set of languages, but we aim to capture phoneme correspondences between languages such as the /{/ – /a/ – /A/ correspondence mentioned above. Our work is, therefore, different from proposals put forward by Deng (1997) who defines a set of universal phonological features to be used for multilingual speech recognition.

The three language-specific vowels discussed above can be grouped together into the metaphoneme |{Aa|, which will be realised as an /{/ in English, an /A/ in Dutch, and an /a/ in German[3]. Tiberius and Cahill (2000) described the vowel metaphonemes for these three languages. In this paper we describe a similar experiment that looked at the consonants of the three languages. The consonants are interestingly different from the vowels for a number of reasons:

- the consonant space is more discrete than the vowel space, so there is less scope for small and non-meaning-bearing distinctions within the consonants;
- the phoneme inventories of the three languages show that, while they have significantly different vowel inventories, their consonant inventories overlap greatly;
- while vowels were considered to occur one per syllable (i.e. long vowels and diphthongs were treated as single vowels), consonants can occur in clusters at either the beginning or end of syllables;
- unlike vowels, consonants can be lost altogether, thus leading to synchronic alternations

[1]This work was supported by ESRC grant no R000223681.

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

[3]We use the notation |xyz| to denote the metaphoneme where x, y and z are normally the sounds for the three languages in the order English, Dutch, and German. However, this is intended only as a mnemonic and does not necessarily imply that these three sounds always occur. Metaphonemes may involve quite complex definitions that are dependent on phonological context as well as just the language in question.

between zero and other segments.

In this paper we present the results of our experiment, first discussing the exact nature of metaphonemes, comparing them with archiphonemes and morphophonemes as used in traditional approaches to morphology as well as the keysymbols used by Fitt (2001) to define cross-accent differences. We then go on to describe the methodology used and the results obtained. Finally we discuss the implications of our findings both on the intended application, i.e. multilingual lexicons, and for other fields such as historical linguistics.

2 Metaphonemes, archiphonemes and keysymbols

The phonemic principle, which has been with us since the end of the nineteenth century, proposes that sets of similar sounds, which can be distinguished by the phonological context in which they occur, can be grouped together to form a single abstract phoneme. The distinct sounds or phones have been defined as allophones of the phoneme, one and only one allophone being permitted to appear in any particular phonological context. The metaphoneme principle states that sets of distinct phonemes that appear in different (but related) languages may be grouped together in a similar way as an abstract metaphoneme, where the conditioning factor is the language in question rather than the phonological context. This is the simplest case, but we also allow phonological conditioning to play a part in the definition of metaphonemes. For example, we may want to say that where English has /s/, German has /S/ if it is in the onset and appears immediately before a /t/ but has /s/ otherwise[4].

Archiphonemes (Trubetzkoy, 1939) are used to generalise over phonemes within a language to represent cases where neutralisations arise in certain contexts. For example, for stops that immediately follow /s/ in English, there is no voicing distinction ('skin', for example, cannot be contrasted with 'sgin')[5]. Trubetzkoy proposed that in such cases we use a different symbol to denote the underspecified or neutralised sound. Similarly, morphophonemes (or systematic phonemes) have been proposed by generative grammarians (Chomsky, 1964) to represent situations where distinctions are neutralised in certain morphological contexts. For example, the voicing of the final consonant of the stem in 'knife' and 'knives' is determined entirely by the presence or absence of the plural suffix.

Although there is a superficial similarity between archi- and morphophonemes and metaphonemes, there are a number of crucial differences. We should note first that both archi- and morphophonemes were introduced as an answer to a problem that we do not actually face – namely the problems of violation of the phonemic principle. It is only if one needs to insist on biuniqueness, invariance and linearity that a solution to the potential problem is needed. In the overall approach to phonology and morphology advocated in the present work these are simply not necessary. We allow lexical entries (or definitions of lexical classes) to specify phonological and morphophonological alternations without being restricted to the phonemic principle. Thus, a phoneme in a language can be a realisation of more than one metaphoneme. The other most obvious difference is that archiphonemes are defined only within a single language, whereas metaphonemes are defined across languages. In terms of the overall theory of morphology, phonology and the lexicon into which metaphonemes were designed to fit, the generalisations represented by metaphonemes come at a different level from archiphonemes.

The keysymbols proposed by Fitt (2001) are much closer to our metaphonemes. The most obvious difference here is that metaphonemes range over languages, while keysymbols are defined across different accents of a single language. However, this apparently significant difference is only sustainable if we maintain that there is a solid definition of what is a language and what is a dialect (or accent). We would maintain that the type of lexicon which represents related languages according to a hierarchical definition of their similarities can be extended very simply to represent distinct dialects of a single language in exactly the same way. However, there are practical differences in the way Fitt's keysymbols and our metaphonemes are employed. Fitt assumes a text-to-speech application in which the same words are to be pronounced, but in different accents. We assume a more general lexicon system, in which we may want to represent differences in whole dialects, not just accents, so that not only the pronunciation will be different. Fitt's system al-

[4]This is still very simplified. See below for a more detailed discussion of this particular metaphoneme.

[5]In fact the realisation of such consonants is somewhere between the voiced and voiceless forms, with minimal actual voicing, but no aspiration that is usually associated with voiceless stops.

lows the definition of a single lexicon which outputs ambiguous strings, including keysymbols, to a speech synthesiser which interprets the keysymbols and disambiguates the pronunciation to get that desired. In the case of metaphonemes, we anticipate a lexical structure which allows lexical entries to be ambiguous as to their pronunciation, but the output of the lexicon as a whole is unambiguous, the metaphonemes being expanded out to their realisation in the different languages (or dialects) as part of the output process from the lexicon.

3 The metaphonemes of Dutch, English and German

In order to define the metaphonemes, we constructed a database of around 800 cognate words from the three languages. The database began with orthographic forms, to which we automatically added the phonological forms from CELEX[6]. We then slightly massaged the database so that leading or trailing schwa syllables were ignored and for most cases just the core root was left for each language. Finally, we analysed the forms into syllabic structures and collated the onsets, peaks and codas for each language[7].

With this information we did two things: first we looked at the absolute correspondences, for clusters and for single consonants, and their frequencies. That is, we considered each grouping of correspondences, such as:[8]

```
st+st+St
str+str+str
nd+nd+nt
m+m+m
k+k+k
```

This gave us both some idea of the likely correspondences and some suggestions as to how phonological context might affect them. We did this for onsets only, codas only and for the two combined. Secondly, we extracted all of the individual consonant correspondences. This had to be done semi-manually as we wanted to ensure that, in cases such as `sk+sx+S` the correspondences came out as `s+s+S`, and `k+x+0` (e.g. 'school', 'school', 'Schule'). From this we derived a set of tables[9] which give, for each consonant in each language, the consonants it can correspond to in the other two languages and how often it does so in our cognate database.

As we expected, there were many cases where the consonants in question were almost always the same across the languages (e.g. `m+m+m`). Also as we expected, the most interesting areas were where one or more languages have different phonological constraints (e.g. /St/ in German onsets vs /st/ in Dutch and English onsets) or where one or more languages have a phoneme that the other(s) do not (e.g. /pf/ in German, /G/ in Dutch).

3.1 Analysis of results

The tables themselves give us a great deal of information, but the whole story can only be gleaned from both sets of data taken together. Let us now consider in detail one small area of the analysis, that covering the consonants /s/, /S/ and /z/. The sounds are obviously related phonologically. /s/, /S/ and /z/ are the only sibilants that occur in all three languages. Figure 1 shows the relevant tables for these sounds starting from English. Just looking at these tables tells us that for /S/ in English, there is just a single metaphoneme worth defining, namely |SsS|, i.e. English /S/ maps to Dutch /s/ and German /S/. The table for /s/, however, shows us rather more interesting things. For English /s/, Dutch has two clear possibilities, /s/ or /z/, while German has three, /S/, /z/ or /s/. To determine how these are related we need to look at the original correspondence database so that we can see if there are any patterns for the possible correspondences. The relevant entries[10] from the first data set for onset only are:

```
31    s+z+z
15    st+st+St
14    S+sx+S
 6    str+str+Str
 5    sw+zw+Sv
 5    sp+sp+Sp
 4    s+s+z
 3    s+z+0
```

[6] Where there were homographs with different pronunciations, the choices between them were made manually.

[7] It should be noted that this process is entirely automatic, and could be applied equally to databases of other cognate languages (e.g. French, Italian, Spanish). Indeed, it would also be possible to construct a database that included for English the cognates from other languages (e.g. French). There will inevitably be gaps in the cognate mappings for any set of languages, a database that maps some English words to one language and other words to another language would be just as acceptable as the database we have worked with to date.

[8] Note that we use the ordering English, Dutch, German throughout.

[9] The full tables for vowels and consonants are available at `http://www.itri.bton.ac.uk/projects/metaphon`.

[10] This is a reduced set for simplicity.

English		Dutch		German	
s	131	s	83	S	39
		z	42	z	37
		t	3	s	35
		w	1	0	9
		l	1	ts	5
		0	1	l	2
				v	1
				x	1
				r	1
				g	1
total	131	total	131	total	131

English		Dutch		German	
z	9	s	5	z	6
		z	4	r	2
				S	1
total	9	total	9	total	9

English		Dutch		German	
S	26	s	21	S	21
		k	1	0	3
		l	1	k	1
		z	1	z	1
		0	1		
		d	1		
total	26	total	26	total	26

Figure 1: Tables for /s/, /z/ and /S/ in English

```
 2    st+st+0
 2    spr+spr+Spr
 2    sl+sl+Sl
 2    sk+sx+S
```

for coda only they are:

```
21    t+t+s
16    st+st+st
11    s+s+s
 5    S+s+S
 4    z+s+z
 3    st+st+0
 3    ks+s+ks
 2    0+s+s
```

We can see that in German, whereas /st/ appears in coda position (corresponding strongly with /st/ in both Dutch and English), in onset position /St/ appears corresponding with /st/ in Dutch and English. Indeed, /s/ followed by a consonant in English and Dutch onsets tends to correspond to /S/ followed by that consonant in German onsets. We could speculate on many possible implications of the clustering of consonants, but in the majority of cases, the absolute correspondences across the languages are so strong that we gain very little by considering phonological context. However, this is clearly a case where phonological context is useful. The tables themselves suggest six possible metaphonemes for English /s/: |ssS|, |ssz|, |sss|, |szS|, |szz| and |szs|. The third of these we can eliminate as it is simply the default case where all languages have the same segment. From the data above, we can see that the metaphoneme |ssS| is likely to be a very useful one, as it occurs in many onset clusters.

The data above, however, allow us to say even more. When we look at the distribution of the /s/ and /S/ in German, it is evident that a metaphoneme that specified that English and Dutch both have /s/ in all contexts while German has /s/ in the coda and /S/ in the onset would capture a much wider generalisation, and cover 74 of the 131 English /s/ cases. This then leaves us with the alternations that involve /z/ in Dutch and German. We therefore propose a metaphoneme |szz|, which is clearly evidenced by the 31 cases of this simple correspondence for onsets above. However, looking more closely again, we can see that this correspondence does not occur at all in the coda, where English /s/ (on its own) corresponds to /s/ in both Dutch and German. This is clearly a result of final consonant devoicing in these two languages, and can be captured by making the metaphoneme defined above phonologically conditioned. Thus, English /s/ corresponds to /z/ in the other two languages in the onset, and to /s/ in the coda.

4 Implications of the results

The intended application of metaphonemes is hierarchically organised multilingual lexicons that permit the sharing of information at all levels (Cahill and Gazdar, 1999), potentially useable for speech recognition or synthesis. The use of metaphonemes allows us to greatly increase the amount of sharing of phonological information across related languages in such a multilingual lexicon. As Tiberius and Cahill (2000) described, using metaphonemes for the vowels alone increased the amount of phonological definitions that could be shared by around

25%. While the use of consonant metaphonemes does not lead to such significant increases in sharing, we estimate that the combined figure rises to around 40%.

Introducing metaphonemes may also be beneficial with respect to the robustness of NLP systems. Knowledge about cross-linguistic commonalities can help to provide grounds for making 'intelligent guesses' when lexical items for a particular language are not present. For example, consider the lexical entry for English 'plough'. We hypothesise a metaphoneme |pppf| (/p/ in English and Dutch, /pf/ in German) as well as |aUu:u:| (/u:/ in Dutch and German, /aU/ in English) and |0xg| (/0/ in English, /x/ in Dutch and /g/ in German)[11]. If we know that the English word 'plough' has the form /plaU/ and that the corresponding Dutch word 'ploeg' has the form /plu:x/, we may predict that the German form would be /pflu:g/. In fact, the German 'Pflug' has the form /pflu:k/, due to the pervasive final consonant devoicing. Thus we can see that in such a case, metaphonemes may help us to predict a form, although the result will not necessarily be fully correct. This example also illustrates the usefulness of phonological conditioning, as we would surely want ultimately to define all consonant correspondences in German and Dutch to take account of the final consonant devoicing process.

Another potential use for metaphonemes is in the field of second language learning, where the typical errors made by learners of a language may be determined by unconscious use of corresponding sounds from their own language.

As well as giving a good indication of possible candidate metaphonemes, the analysis we performed also gave us other information about the three languages which is potentially of interest to historical linguists. The analysis we did involved matching the corresponding segments in forms which are originally from identical roots. Thus we might expect that the data can give us clues about how the languages have changed and diverged. For example, a zero in a possible consonant position in one language suggests that that language has lost a segment where (at least one of) the other languages still have one. Looking at which segments are found in such positions gives us a clue as to which segments are most likely to be lost in language change (at least in these languages). Indeed, it transpires that the highest ranked segments in these positions are, as one would expect, mostly approximants, liquids and glides (/r/, /w/, /l/ etc.). Also interesting is that of the stop consonants, the most likely to be lost in all three languages are the velar consonants /k/ and /g/. Another interesting result from this examination is that Dutch is apparently less likely to have zeros than German, while English is much more likely to have zeros than either of the other two languages. (41 for Dutch compared to 147 in German and 268 in English).

[11] All of these metaphonemes are predictable from the full correspondence tables.

5 Conclusions

In this paper, we have presented the theory of metaphonemes. We have illustrated our theory with the definition of cross-linguistic phoneme correspondences for English, Dutch, and German. We believe that our work has interesting benefits for speech applications. The information that can be specified in metaphonemes can be used to tune speech applications to closely related languages in a similar way that Fitt's (2001) keysymbols are used to model different accents. The next steps in this research will involve fully integrating the metaphonemes into a multilingual lexicon to enable testing on a speech synthesis system.

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

Chomsky, N. 1964. *Current Issues in Linguistic Theory*, Mouton, The Hague.

Deng, L. 1997. "Integrated-Multilingual Speech Recognition using Universal Phonological Features in a Functional Speech Production Model", In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol II Speech Processing*, Munich. Germany. pp.1007-1010.

Fitt, S. 2001. "Morphological Approaches for an English Pronunciation Lexicon." In *Proceedings of Eurospeech 2001*. Aalborg, Denmark.

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

Tiberius, C. and L.J. Cahill. 2000. "Incorporating Metaphonemes in a Multilingual Lexicon." In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbruecken, Germany, pp. 1126-1130.

Trubetzkoy, N. 1939. *Grundzuge der Phonologie*, Vandenhoeck and Ruprecht, Gottingen.

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.