Phonetic annotation of a non-native speech corpus

Patrizia Bonaventura[1]*, Peter Howarth*[2] *and Wolfgang Menzel*[3]

[1]Conversa, Redmond, WA, [2]University of Leeds, [3]University of Hamburg

PatriziaBonaventura@conversa.com, p.a.howarth@leeds.ac.uk, menzel@informatik.uni-hamburg.de

Abstract

Annotating non-native speech on a phonetic level is an extremely labour-intensive task and therefore requires a proper balance between the expected benefit and the resources needed. This paper reports on the experience gained when collecting and annotating a corpus of English sentences recorded by students with Italian and German as their mother tongue. The annotated data were used intensively during the development phase of a language learning tool, which allows automatic diagnosis of pronunciation errors and gives corrective feedback to the learner. Suggestions for further improvement in the annotation procedure will be presented, based on the experience acquired in creating the corpus.

1. Introduction

Using advanced speech technology to provide a foreign language student with corrective feedback on possible mispronunciations seems a promising step forward in the development of a new generation of computer-based language learning systems. Based on dedicated components for the diagnosis of phone and stress errors such a system is not only able to identify a student's mistakes, but moreover can present additional material to focus on the problem, is able to provide guidance for improvement and might even offer specifically tailored opportunities for individualised practice (Herron et al., 1999, Menzel et al., 2000). Since the underlying technology relies heavily on the availability of a substantial amount of carefully annotated non-native speech data, great care has to be taken in generating the necessary material, which is needed for the development of the individual components, the fine tuning of the overall system and the evaluation of its performance.

Unfortunately, the phonetic annotation of speech data is a cumbersome and labour-intensive (and therefore costly) procedure, which involves developers finding an acceptable compromise between the technical requirements of the system development process and the available financial resources. One of the major factors, which makes any phonetic transcription such an expensive task, is the inter- and intra-personal variability of human annotators. This problem has already been noticed in a monolingual context (Eisen et al. 1992). It becomes even more serious when non-native speech is considered, which naturally can be described both from the perspective of the student's mother tongue (L1) and of the language to be learnt (L2). This not only introduces an additional source of disagreement but also makes the definition of an appropriate symbol inventory even more problematic.

The corpus that has been collected consists of a set of 164 sentences, containing approximately 1100 words. They have been selected in order to be used in different exercises, to provide German and Italian learners with the chance to practise problem phones, weak forms, consonant clusters, stress patterns and contrastive intonation of the foreign language (Menzel et al., 1999). All sentences have been recorded by 46 speakers (23 German and 23 Italian). As far as possible speakers were selected with an intermediate level of proficiency in English.

2. The annotation procedure

All speech files were first automatically annotated at the phone level with canonical pronunciations aligned by look-up from the dictionary of a triphone-based HMM recogniser (Morton et al., 1999). Primary stress was also annotated by look-up in the CELEX lexical database for English (Pipenbrock, 1993).

In addition to the phone level the corpus was also annotated at the word and stress levels. A primary stress marker was placed in the phonetic string. Unstressed syllables were marked, but monosyllabic words would retain the stress reported in the canonical annotation.

The symbols available for annotation closely correspond to the phonemes of the underlying recogniser: centralised diphthongs [ɪə], [eə], [ʊə] (e.g. in 'fear', 'stair', 'poor'), were represented by single symbols. Table 1 in the Appendix gives the inventory of British English phones used both in the recogniser and for the transcriptions, as well as the mapping of these phones to the most similar Italian and German ones. The Italian and German phones marked in bold are those chosen to be modelled in the recogniser as non-native variants, based on Italian and German native speech material, because they represented phones too different from the English ones. Among German consonants, for example, [r] represents all /r/ allophones in German, i.e. the central low vowel [ɐ], occurring very often after vowels, and the uvular trill [ʀ], occurring in initial position before vowels, intervocalic and between consonant and vowel. Italian [r], on the other hand, represents the dental trill [r]. All these phones are very different from the most common realisation of English /r/, i.e. the post-alveolar voiced flap [ɹ]. Among the vowels, Italian [e], for example, can stand either for the English-like [e], or for the Italian-like realisations [ei], [ɛ], [i].

Annotation

Five annotators were recruited in addition to a member of the project team from the University of Leeds, who acted as co-ordinator. The six were all native speakers of British English and trained in phonetics. The five recruits were undergraduate or postgraduate students in the Department of Linguistics and Phonetics. One result of the degree of training most of them had was their difficulty in adjusting to the rather reduced set of transcribed phones available to them in the annotation scheme and their unfamiliarity with the recogniser-specific symbol set (eg ‘ax’ for schwa). In cases of clear non-native phone interference (for example the Italian trilled [r]), annotators were able to choose a phone from alternative non-English phone sets, but they were encouraged, where possible, to select the ‘closest’ match from the UK English phone set.

The complete annotation task was divided into 'sessions'; a session comprising 255 sentences recorded by one speaker (the total included utterances to be used for adaptation). The data from 23 Italian and 23 German speakers produced the 46 main sessions. In order to check consistency in annotation, five sets of pseudo-speaker data were created, each one also made up of 255 utterances, composed of arbitrarily selected utterances from each speaker. All six annotators processed the same five 'pseudo' speakers, amounting to a further 30 sessions. The total number of sessions was therefore 76, and with four files associated with each utterance, the total number of files to be managed amounted to over 75,000.

An initial training session was given to all annotators, with a set of written instructions. It soon became apparent that the preparation was not sufficiently detailed, since once the annotation began in earnest numerous queries arose that hadn't been envisaged. One of the most frequent related to the way in which phones were represented in the look-up dictionary. For example, the vowel in the word 'the' seemed to exist in three canonical forms ([ə], [i:] and [ʌ]), and the one selected by the automatic transcription process for any one instance appeared random. Many of the alterations to the data made by the annotators involved corrections of this kind.

The 76 sessions were distributed to the annotators one at a time. The intention was to give each one a balanced set of male and female, Italian and German speakers. However, this was hard to maintain, as the annotators differed considerably in the speed of their work, and the pressure of time meant that the faster workers received more sessions. The six annotators each completed between 4 and 12 sessions of the 46 real sessions.

The first stage of the annotation process required the identification of any errors at word level in the utterances to be used for the exercise material. These files were then re-aligned and sent back to the annotators for them to continue checking the stress and phone levels. The session was complete once all files had been checked in this way. After this stage there were a few files returned to annotators for clarification, especially when some non-standard sequence of symbols had been used (mostly related to the use of non-UK phones such as German or Italian R).

The annotation process was completed in four months. Each session is estimated to have taken between 5 and 7 hours, producing a total figure of 400-500 hours.

The annotation tool

One of the project partners, Entropic Cambridge Research Laboratory Ltd., provided a PC-based tool, *WavEd*, with which the waveform can be annotated with errors. The tool displays the time-amplitude speech waveform, and aligned reference transcriptions at each of the three levels (word, phone and stressed syllable), as shown in Figure 1. The waveform can alternatively be displayed as a spectrogram. The audio file for a whole utterance can be played back, or the annotator can highlight a section to listen to. When viewing a whole utterance, it is usually not possible to display all the phone and stress labels, so a zoom facility is provided for them to be viewed more easily.

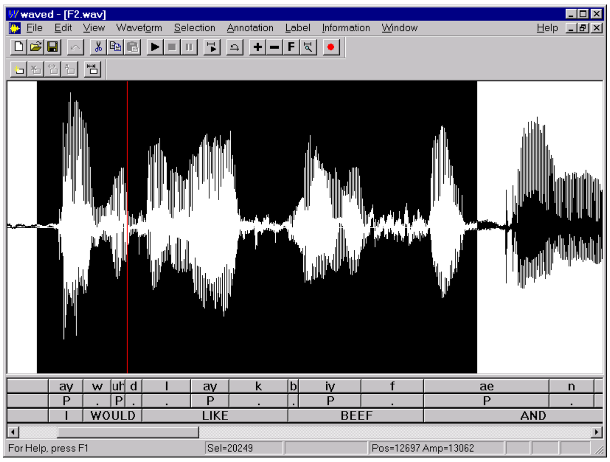

Figure 1: The *WavEd* annotation tool

The annotator can click on a label at any of the levels of annotation below the waveform, and edit it to indicate that the label is incorrect, i.e. the speaker has said something other than the canonical pronunciation provided in the reference transcription. Such changes consist of deletions, insertions and substitutions at each of the three levels of units. Deletions are shown by replacing a unit's label with a zero. Insertions are indicated by appending a hyphen and the extra unit label to the leftmost neighbouring unit, or adding the extra unit label and a hyphen to the rightmost neighbour. Substitution is shown by simply replacing the label with the observed one. Examples can be see in Table 2:

Error Type	Reference Label	Edited Label	Example
Deletion	h	0	dropped 'h' in *how*
Insertion	f	f-ax	word-final schwa insertion on *beef*
Sub-stitution	uh	uw	*book* rhyming with *boot*

Table 2: Examples of marking-up conventions

The *WavEd* tool does allow the boundaries between units to be moved, but the annotators were instructed not to change the boundaries, even if they were wrong, since this would disturb the integrity of the alignment between the three annotation levels.

Examples of pronunciation errors identified

Examples of pronunciation errors at each level, subdivided between German and Italian native speakers are given below, with an indication of whether these are expected (owing to L1 interference and attested in the EFL literature) or unpredictable/idiosyncratic.

Word level (not systematic or easily predictable)

Italian

photographic → photography
than/then → that
deserted → desert (phone level error?)
like to → to like
+ the
- to

German

not be → be not
the → a
month → week
of → about
+ more
- in

Stress level (largely as predicted)

Italian

ʹphotographic
ʹconvict / conʹvict
ʹcomponents

German

ʹreport
ʹtelevision
contrast / contrast

Phone level (as predicted + idiosyncratic)

Italian

said: [e] → [eɪ]
ticket: [ɪ] → [i:]
biological: [aɪ] → [i:], [ɒ] → [əʊ],
[ɪ] → [i:], [ə] → [æ]
shoulder: [əʊ] → [u:]
honest: + [h]
thin: [θ] → [t]
sleep: [s] → [z]
ginger: [dʒ] → [g] (2x)
singer: [ŋ] → [n g]
sheep: + [ə]
bait: - [t]

German

pneumatic: + [p], - [j], [u:] → [ɔɪ]
ʹproduce: [ɒ] → [əʊ]
said: [s] → [z]

visa: [v] → [w]
weekend: [w] → [v]
the: [ð] → [d]
biscuit: + [w]
thumb: + [b]
finger: - [g]
dessert: - [t]
cupboard: + [p], [ə] → [ɔ:]

Annotator reliability and accuracy

The transcriptions of the pseudo-speaker data (processed by all six annotators) were analysed and the following results found:

47% inter-annotator agreement on the quality of the error and 67% agreement on error localisation. Figure 2 shows the correlation between different annotators and with themselves in terms of their hit rate (i.e. the percentage of errors identified by annotator A, which were also found by annotator B). Here, the smaller bars represent strong hits (i.e. both annotators agree on the kind of error) whereas the taller ones indicate weak hits (i.e. the annotators only agree on the place of the error). Figure 3 gives the false alarm rate, i.e. the percentage of errors one annotator found but the other did not.

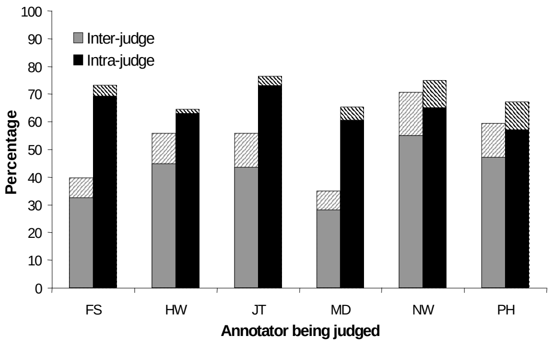

Figure 2: Hit rate across annotators

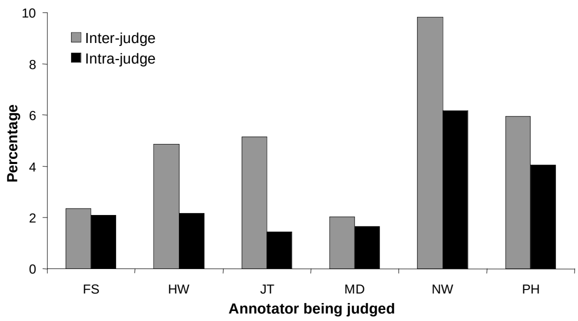

Figure 3: False alarm rate across annotators

Transcriptions have been further double-checked by a native Italian speaker with a good command of English and German. Judgements often turned out to be biased

a) by the influence of the mother-tongue of the transcriber, and

b) by an annotator's insufficient proficiency in the mother tongue of the non-native speaker.

Examples of the former effect are e.g.

1) the transcription of the final orthographic sequence 'er' in the word 'photographer' by the English annotator, as [ɜː] (as in the canonical pronunciation [f ə 't ɒ g r ə f **ɜː**]), whereas the actual German mispronunciation was [ɐ], according to the German spelling-to-sound mapping for final 'er'. The German phone [ɐ], which did not belong to the available inventory, would have been better represented by the articulatory and acoustically closest English vowel [ʌ],

2) the transcription of the initial orthographic sequence 'ps' in the word 'psychology' by the English annotator, as [s] (like in the canonical pronunciation [s ː ɪ 'k ɒ l ə dʒ i]), whereas the actual German mispronunciation was [p s], according to the German spelling-to-sound mapping for initial 'ps',

3) the transcription of the initial orthographic sequence 'pn' in the word 'pneumatic' by the English annotator, as [n] (like in the canonical pronunciation [n j uː 'm æ t ɪ k]), whereas the actual German mispronunciation was [p n], according to the German spelling-to-sound mapping for initial 'pn',, or

4) the transcription of the initial orthographic 'w' in the word 'would' by the English annotator, as [h w] (like in the very common non-standard realisation of [w ʊ d] as [h w ʊ d]), whereas the actual Italian mispronunciation was [h ʊ d], according to the simple pronunciation rule for initial 'w' before vowel, commonly taught in the Italian schools.

Examples of annotation problems caused by insufficient knowledge of the student's mother tongue are:

1) the transcription of the final sequence 'arb' in the word 'rhubarb' (canon. pron. ['r uː b **ɑː b**]) as ['r uː b **æ r b**] by the English annotator, whereas the German rendition was ['r uː b **ɑː p**], reflecting a carry-over of the spelling-to-sound rules and of the phonological regularities of the speaker's L1, i.e. pronunciation of 'ar' before consonant as [ɑː], and the unvoiced realisation of final voiced stops,

2) the transcription of the orthographic sequence 'ch' in the word 'psychology' by the English annotator, as [k] (like in the canonical pronunciation [s ː ɪ 'k ɒ l ə dʒ i]), whereas the actual German mispronunciation was [ç], according to the German spelling-to-sound mapping for 'ch' after front

vowel; the German phone [ç], which did not belong to the available inventory, would have been better represented by the articulatory and acoustically closest English consonant [ʃ]

3) the transcription of the orthographic sequence 'eu' in the word 'pneumatic' by the English annotator, as [əʊ] (canonical pronunciation is: [n j u: 'm æ t ɪ k]), whereas the actual German mispronunciation was [oɪ], according to the German spelling-to-sound mapping for 'eu'; the German diphthong [oɪ], could have been better represented by the corresponding English [oɪ], or

4) the transcription of the orthographic sequence 'dd' in the word 'udder' by the English annotator, as [d] (like in the canonical pronunciation: ['ː d ə]), whereas the actual Italian mispronunciation was [dː d], according to the Italian spelling-to-sound mapping for double consonants, which are realized as geminates (Canepari, 1980); the Italian long consonant [dː d], could have been better represented by a reduplication of the corresponding English [d] as [d d].

Furthermore, some transcription inaccuracies resulted from the lack of phonetic details in the annotation scheme for non-native pronunciation phenomena, e.g. partial de-voicing of labiodental fricative [v] after the unvoiced velar stop [k] in the word 'quite' (canonical pronunciation ['k w aɪ t]), produced by Germans, could not be represented by the appropriate diacritics (i.e. IPA [̥] for devoiced phones, hence [kv̥aɪt]). This and many other potentially useful options were lacking in the adopted set of symbols.

It is very common in non-native speech to find intermediate phones between the canonical IPA sounds, which can mislead the perception of even an expert transcriber: an accurate transcription of such sounds is probably possible only after experimental measurements. Therefore, it would be preferable for all annotators to have a good command of both the target language to be taught and the L1 of the non-native student.

Since the inter-annotator agreement was rather low, a larger number of annotators would be desirable (at least two natives of the language to teach, to transcribe the corpus initially and at least twice as many annotators subsequently involved in correcting just the transcriptions of the mistakes). However, such a highly redundant approach would render the transcription process impractically expensive. For the purposes of economy, just two annotators could be used: should there be a disagreement about the transcription of a mistake, further checks would be necessary, to be carried out both individually and by discussion or by other measurements.

Further problems encountered in annotation

1) There were insufficient phones available for annotators to apply. This was particularly apparent with diphthongs such as /eə/ in 'where'.
2) The assumption of rhotic pronunciation (because of the US origin of the annotation scheme) meant that a lot of words were transcribed with postvocalic R, when this would not be the target pronunciation in UK English. Annotators were told to ignore this, though this introduced some confusion.
3) There were occasions when it was not clear whether an error should be classed as occurring at phone or word level. For example, saying 'comes' in place of 'come' could be interpreted either as a mis-reading of the text on screen or as the result of phonological habit of the speaker. The decision requires a consideration of context and an understanding of non-native speech production, and therefore some subjective judgement.
4) A particular difficulty in analysing the Italian data was determining whether or not a final schwa has been added. It soon became clear that there was a cline from zero added to a clear extra syllable on a word such sheep (ʃi:pə). The problem arose in differentiating between an added schwa and an strongly aspirated final stop consonant.
5) It is probable that the six annotators, necessarily working independently, developed individual ways of reaching compromises in their judgements.
6) It was unfortunate that the results of the inter-annotator reliability tests could not be fed back into the annotation process, as this might have been useful for modifying and standardising judgements. Overall the pressure of time meant that it was difficult to review the procedures systematically and learn from experience.

3. Recommendations for annotation of non-native speech

Drawing on experience derived from previous dubious transcriptions, some suggestions for the annotation procedure for non-native speech can be put forward:

1) A sufficiently rich set of phonetic symbols should be used, to allow the representation of all phones of the target language, including some additional diacritics to represent non-phonological sounds produced by the non-native speakers, intermediate between the L1 phonemes and the target L2 phonemes.

2) The level of phonological/phonetic competence of the non-native speakers should be homogeneous within the annotation task assigned to each annotator, in order to limit the number of possible realisations of a given L2 phone: speakers with different levels of knowledge of L2 would presumably produce mistakes more inconsistently than speakers at any one level of performance. At least three levels should be considered (high, intermediate and low). A reliable method of determining the level of performance of the speakers would be to administer an articulation test, e.g. the "Arizona Articulation Proficiency Scale" (Barker Fudala, 1974) adapted for testing adult non-native speakers' proficiency.

3) Once the level of competence of the speakers has been controlled, the corpus should be annotated via the following steps:

   a) A trained phonetician, native speaker of the target language to be taught, possibly with good knowledge of the L1 of the non-native speaker, should first review the canonical transcriptions, marking the portions of the signal that he perceives as being incorrectly pronounced, and transcribe the mistakes. If the annotator is not sure about a transcription, the mispronunciation can be marked as uncertain.

   b) A trained phonetician, native speaker of the L1 of the non-native student, with good knowledge of English, will listen again to the portions of the speech material previously marked as mistakes. This annotator can confirm every transcription by the previous transcriber. If he would like to correct a transcription that the native English annotator has considered certain, he can do so after an experimental check (e.g. measurement on spectrograms). If, however, he does not agree on a transcription previously marked as uncertain, he can apply any changes without further testing, as long as he is positive about the transcription; otherwise, an experimental check is necessary.

4. Conclusions

The phonetic annotation of non-native speech corpora constitutes a much more expensive procedure than the annotation of a speech corpus in a given natural language. Ehzani and Knodt (1998) estimated a price of up to $1 per phone for non-native data. In the ISLE project the rate is estimated to be 0.6 Euros per phone, suggesting a relatively economical data resource which has proved highly versatile for the development of a pronunciation training system. The recommendations for transcription outlined above might further improve the quality of the end result within affordable limits.

The corpus is available for non-commercial purposes through the European Language Resources Distribution Agency (ELRA).

5. Acknowledgements

This research has been supported by the European Commission under the 4th Framework of the Telematics Application Programme (Language Engineering Project LE4 8353).

References

Barker Fudala, J. (1974) Arizona Articulation Proficiency Scale, Western Psychological Services, Los Angeles.

Canepari, L. (1980) Introduzione alla fonetica, Torino, Einaudi.

Ehzani, F. Knodt, E. (1998) Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. in: Language Learning & Technology, vol. 2, no. 1, p. 45-60.

Eisen, B., Tillmann, H. G., Draxler, Ch. (1992) Consistency of Judgements in Manual Labelling of Phonetic Segments: The Distinction between Clear and Unclear Cases, Proc. Int. Conf. On Spoken Language Processing ICSLP'92, p. 871-874.

Herron, D., Menzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Morton, R., Schmidt, J. A. (1999) Automatic Localization and Diagnosis of Pronunciation Errors for Second Language Learners of English. Proc. 6th European Conference on Speech Communication and Technology, Eurospeech '99, Budapest, vol. 2, p.855-858.

Menzel, W., Herron, D., Bonaventura, P., Morton, R. (2000) Automatic detection and correction of non-native English pronunciation. This volume.

Morton, R., Whitehouse, D., Ollason, D. (1999) "The IHAPI book". Entropic Cambridge Research Laboratory, Ltd.

Piepenbrock, Richard. (1993) A longer term view on the interaction between lexicons and text corpora in language investigation. In Souter, C. and E. Atwell (eds.) Corpus-Based Computational Linguistics. Amsterdam: Rodopi, pp. 59-70.

Appendix

Table 1. Inventory of English phones (in IPA and SAMPA symbols), and the mapping, used in the recogniser, with the most similar Italian and German phones (in bold are the Italian and German phones which were modelled independently from the English phones in the recogniser)

English phones (IPA)	Entropic for ISLE	SAMPA	Example English word	Italian phones (IPA)	Example Italian word	German phones (IPA)	Example German word
p	p	p	pin	p	papa	p	Paar
b	b	b	bin	b	bene	b	Ball
t	t	t	tin	t	tata	t	Tafel
--	t s	--	--	ts	stazione	--	--
d	d	d	din	d	dare	d	Deutsch
--	d z	--	--	dz	zero	--	--
k	k	k	kin	k	con	k	Kind
g	g	g	give	g	gatto	g	Gern
f	f	f	fan	f	fiume	f	Fern
v	v	v	van	v	va	v	Wer
θ	th	T	thin	--	--	--	--
ð	dh	D	this	--	--	--	--
s	s	s	sin	s	sole	s	fassen
z	z	z	zing	z	asma	z	singen
ʃ	sh	S	shin	ʃ	lascia	ʃ	Stein
ʒ	zh	Z	measure	--	--	ʒ	Genieren
h	hh	h	hit	--	--	h	Hand
tʃ	ch	tS	chin	tʃ	cera	x	Loch
dʒ	jh	dZ	gin	dʒ	gelido	dʒ	Joystick
m	m	m	mock	m	mamma	m	Matt
n	n	n	knock	n	nonna	n	Nest
--	n y	--	--	ɲ	gnocchi	--	--
ŋ	ng	N	thing	--	--	ŋ	lang
l	l	l	long	l	lardo	l	Links
--	l y	--	--	ʎ	gli	--	--
r	r	r	wrong	**r**	caro	**r**	rennen

English phones (IPA)	**Entropic for ISLE**	**SAMPA**	**Example English word**	**Italian phones (IPA)**	**Example Italian word**	**German phones (IPA)**	**Example German word**
j	y	j	yacht	j	paio	j	Million
w	w	w	wasp	w	uovo	--	--
ɪ --	ih --	I --	pit --	-- --	-- --	ɪ y	Kiste System
e --	eh --	e --	pet --	-- --	-- --	e ɛ:	wem Affäre
æ	ae	{	Pat	**ɛ**	bene	--	--
ə --	ax --	@ --	another --	-- --	-- --	ə œ	mache Hölle
ɒ	oh	Q	pot	--	--	o	Offen
ʌ	ah	V	cut	--	--	a	Kampf
ɜ:	er	3:	furs	--	--	--	--
ʊ --	uh --	U --	put --	-- --	-- --	**ʊ** œ:	Mutter Flösse
i:	iy	i:	ease	i:	io	i:	Ziel
u: --	uw --	u: --	loose --	**u:** **--**	pure --	u: y:	Mus Kübel
ɔ:	ao	O:	cause	**ɔ**	cuore	--	--
ɑ:	aa	A:	stars	a	gatto	a:	Kahn
əʊ	ow	@U	nose	**o**	canto	o:	Ofen
ɪə	--	I@	fears	--	--	--	--
eə	--	e@	stairs	--	--	--	--
ʊə	--	U@	poor	--	--	--	--
eɪ	ey	eI	raise	**e**	parte	--	--
aɪ	ay	aI	rise	--	--	aɪ	weit
aʊ	aw	aU	rouse	--	--	aʊ	Haus
ɔɪ	oy	OI	noise	--	--	ɔy	freut