Transcribing names - a multiple choice task: Mistakes, pitfalls and escape routes.

Andreas Mengel, Technische Universität Berlin

In this paper I will address the complex variety of mistakes and problems that occur on the long way to our pronunciation dictionary. I will also provide possible solutions.

1. Source data

The data we are given by the PTT's can have spelling errors. In christian names and surnames these spelling errors are nearly undetectable. Of course some names can be excluded e.g. if they do not have any vowels. Generally, misspelled names can only be detected if the data include frequency data: Thus names occuring more than five times for example can be considered correct. But even this heuristic can only give a hint: names of larger towns for example appear more often and thus can be misspelled the same way more often as well. However, this is the only plausible way since even likely names - those that look like names but are no such (like **D**erlin) - can be misspelled ones. Thus another possiblitiy is to use other additional sources - if available.
A completely different problem is that the data are not representative in two respects: age and gender. The first reason is that in principle the names of men (husbands) appear as entry in the PTT data. The second reason is that cildren normally do not have telephones and thus their names are not listed.

Another problem concerning the source data is, that individual orthographical systems are more or less adapted to only one sound system. This has the one effect that for the transcription of foreign names that are originally written in the latin alphabet different LTS rules have to be applied. The second effect is the following one: There are a lot of foreign names that are usually not written with the latin letter system and the pronunciation of which does not fit into the national language sound system of another country. Thus the transcription by means of the national orthography becomes quite complicated and will vary from time to time. Consequently it will become very difficult to produce adequate transcriptions for some foreign names.

2. Transcription

It is very difficult to check automatically if the phonetic entry belonging to an orthographic entry is the transcript of a possible pronunciation of the name. Because the rules themselves are the most evolved machine no other machine can control better the output. That's why [Èpa.pa] is as correct as [Stat] for <Stadt>.
The only means to increase automatic certainty is to have another independent machine check the data as well: These identical transcriptions being produced by both the rule engine and another machine can be taken as valid. This is another reason why we should use neural nets as well. We should't use them only to generate new pronunciations but also to verify our old ones. This is even possible if they are trained by data with a certain percentage of errors if only they are numerous.

3. Phonetic entry

Another possibility to control the phonetic entries is to define a structure of possible phonetic words. This structure should of course contain all features the entries contain themselves. The grammar written for a parser also should not only control the correctness of the symbolset used but also the sound combinations, intonation patterns and the like. The more explicit the grammar for the parser, the closer the net for mistakes. This control procedure and the one mentioned before take into account that man is overtaxed when controlling large amounts of data.

4. Consistency of transcriptons

It is also important to guarantee that equal phonetic patterns are transcribed the same way all over the transcriptions. Therefore lists for the transcribers have to be provided, morpheme lexicons have to be made and heuristic methods like different sortings - alphabetically backward etc. - can help.
A problem that also needs to be taken into consideration is whether variation of pronunciation is caused by dialect or by accent. If variation is produced by dialect and is as such of rather lexical origin and rather unpredictable, it should be transcribed. On the other hand, if variation is produced by predictable accent causing slight vowel-shift or lengthening of sounds that can be perceived in any pronunciation of speakers of a region, one should neglect this variation and standardize.

5. Consistency of lexicon

In order to obtain one dictionary instead of various different ones it has to be made sure, that the same alphabet and the same transcription depth is provided. Therefor it seems indispensible to frequently exchange the outpout data and check them for serious and problematic deviations. The earlier deviations are detected the less will be the effort necessary for changes.

6. Comparability

Another fact to be taken into account and possibly conflicting with the one mentioned before is the comparability of the dictionary with other existing ones. This does not only result in a broader acceptance but also increases reusability of the data produced. Obviously, differences can be permitted if the quality is not affected or only affected in terms of being better, e.g. by providing more information (e.g. syllable boundaries, different stress levels).

7. Informants

The most serious problem is the provision of not only the possible and likely but also of those pronunciations that are produced by the bearers of names and those who frequently use the names. One will have to ask informants. Ways of obtaining these data may be various. But here we have to distinguish between: (1) names where it is transparent to the reader that pronunciation is difficult because of their marked structure; and (2) names that seem quite normal but are pronounced very differently.

As the transcriber has no chance to perceive the abnormality of the latter, one way of obtaining the correct pronunciations is simply to ask people for names that are always mispronounced or misspelled.
It also seems quite clear that those people who have unusual names in one or another respect can be expected to be used to a much wider variety of pronunciations than those whith rather common names: Of course this is no excuse but already the provision of **one** possible pronunciation per name is quite difficult and a group of ten people or less in a country cannot own the competence of the whole population.

8. Pronunciation

For foreign names and unusual native names alike a total of three different transcriptions or realisations is possible for one name. These four possibilities result from two times two different levels of competence in reading and pronounciation minus one, that is implied by another.
Reading competence: within a population there are people who own the competence of recognizing and identifying foreign orthographic name structures and there are people who are innocent in this respect. **Pronouncing competence:** on the other hand there are people who are able to pronounce the sound structures of foreign names because they learned them or their sound-system is comparable to that of the foreign language. Others cannot pronounce foreign sound structures at all. This would give a total of four combinations for foreign names. One can be ommitted immediately because it has no practical relevance: an incompetent listener who is able to produce the correct pronunciation; three remain.

Case (a) is the the native speaker who does not recognize the orthographic pattern as foreign because of lack of knowledge or because it is equal to a native one. He thus cannot produce the correct pronunciation: The normal tourist in a foreign country.
Case (b) is perhaps the most common one: a person who recognizes the pattern as foreign but who cannot pronounce it as he should and thus adapts it to his native phonetic pattern: The culturally interested tourist in a foreign country.
Case (c) is the educated foreign expert who recognizes the pattern as foreign and pronounces it adequately: The German English-teacher in London.

	native reader		occurence	pronunciation of <Timothy> ["tɪ.m«.θi]
	identification by reading	correct pronounciation		
a	-	-	innocent reader/pronouncer	["ti:.mo.ti]
b		-	educated bad pronouncer	["tɪ.m«.si]
c			educated reader and good pronouncer	["tɪ.m«.θi]

Table 1: Possible pronunciations of foreign names correlated to reasons and competences of speakers.