A corpus-based methodology for comparing and evaluating different accents

Martin Weisser
Department of Linguistics and Modern English Language
Lancaster University

1. Learner corpora and their uses

The past few years have seen a growing interest in the development of learner corpora (cf. Granger, 1998). However, so far the main emphasis for this kind of corpus has been on collecting and marking up written texts produced by native and non-native speakers. This was mainly done in order to be able to determine where the most common problems and differences in the usage of written language, both as far as grammar and lexis are concerned, lie for the learners. Learner corpora of spoken language are still relatively rare and even if they do exist, often consist of written orthographic transcriptions, possibly minimally enriched with symbols to indicate lengthening, etc. only. Attempts at producing learner data relating to pronunciation have so far, to my knowledge, been restricted to setting up a flatfile database documenting the realisations of school children, as in the Austrian Learner's Database (Wieden/Nemser, 1991b), and to parts of the ISLE Project[1].

1.1. Corpora vs. LE speech databases

As pointed out above, learner corpora are so far mainly being used to investigate learner behaviour with regard to grammar and lexis, but very little attention has been paid to issues of pronunciation. Phonetic transcriptions of spoken language on a corpus scale are generally the domain of Language Engineering (LE), where the purpose has mainly been to collect speech data that may be used for applications such as speech recognition, speech synthesis, telebanking applications or spoken dialogue systems. However, the material recorded for these purposes is in many cases severely limited with respect to the particular requirements of the application it is being collected for, e.g. series of numbers or certain keywords needed for communicating with the application.

The first notable exception to this kind of corpus is the aforementioned Austrian Learner Database, which documents the pronunciation of Austrian schoolchildren of varying age-groups and from different regions. The purpose of this database, amongst others, was to establish which particular problems in pronunciation for theses learners may be introduced by their L1. However, this database had some serious drawbacks and flaws due of its design and implementation, some of which will be discussed with the relevant topics further below. The second exception is the material collected for the EU-funded ISLE project, which aims to provide an architecture for incorporating speech recognition technology into language-learning software products.

1.2. Transcription, encoding and fonts

Due to the relative high use of UNIX-based workstations in LE, transcriptions in LE speech databases are often based on ASCII representation schemes of the IPA character sets, such as SAMPA (Wells et al., 1992) or use other mapping algorithms, which work well for the computer, but often make it difficult for the human reader to understand the transcriptions, especially if the latter contain many diacritics. The transcription and encoding system used in the Austrian database is even more complex since it not only uses numerical codes to represent the transcriptions themselves, but also includes codes that represent information about the relative closeness of the pronunciations to RP and any deviations from the implicit norms underlying the analysis. For example, one of the representations of the word *rubber* given in Wieden/Nemser, 1991b (p. 354) is "000045400002000010000000050020000". Coding schemes such as this may make it easy to conduct relatively exact statistical comparisons containing minute details of pronunciation, but are extremely difficult to handle and may lead to results biased more towards the quantitative than qualitative side. To facilitate research of a more qualitative nature, the use of standard IPA (True-type) fonts is thus preferable.

[1] http://nats-www.informatik.uni-hamburg.de/~isle/speech_text.html

2. Spoken language models applied to learner English

2.1. RP vs. native speaker corpus-based reference models

When learners of English are assessed or tested, this is usually done against implicit reference models such as RP (Received Pronunciation) for British English or GenAm (General American) for American English because these represent the standardised varieties the learners have supposedly been taught. Furthermore, learners are often expected to enunciate far more clearly and possess a far higher rhetorical skill than native speakers of different varieties of English. Not only is this unrealistic and an unfair practise towards the learners, but it is also, at least as far as British English is concerned, an unrepresentative model of the language as RP is only spoken by about 3 % of the overall population (Hughes/Trudgill, 1997: 3), and even then in different forms (Wells, 1982: 279-301). However, so far no attempts at creating more realistic models of native speakers English have been made that could serve as an adequate basis for comparison between native and non-native speakers.

3. Creating speaker/learner population models

The methodology I propose here represents a synthesis of ideas from corpus linguistics – especially learner corpora – and LE. Creating a kind of reference model from the native speaker realisations in the first instance makes it possible to compare native and non-native speaker data by establishing an adequate basis for comparison, rather than relying on the more abstract established teaching models, such as RP. Such a model would highlight tendencies and patterns in native speaker variation against which one can then evaluate the performance of non-native speakers realistically, and in turn identify problem areas that suggest possible changes in teaching methodology and practice. Apart from comparing native and non-native speakers in this way, there are also other implications and usages of such a methodology, which are discussed towards the end of this paper.

3.1. What needs to be stored?

In order to create suitable models for analysing the spoken language of different speaker populations a number of different types of information need to be stored. Just like in corpora of written language, it is important to store orthographic representations of the spoken material and as far as possible enrich these by at least incorporating morphosyntactic information in the form of grammatical tags[2]. In my implementation, the orthographic material, i.e. a dialogue created and read by 7 native and 10 non-native speakers, is stored word-by-word in a table and information about the word-classes, i.e. grammatical tags, in another.

More than for purposes of handling written language, this type of information needs to be complemented by detailed information about the speakers, such as age, sex, place of birth, etc. because these factors may have a strong influence on the speech behaviour. For non-native speakers it is also important to record details about their exposure to the target language, in order to be able to establish ‘proficiency’ levels.

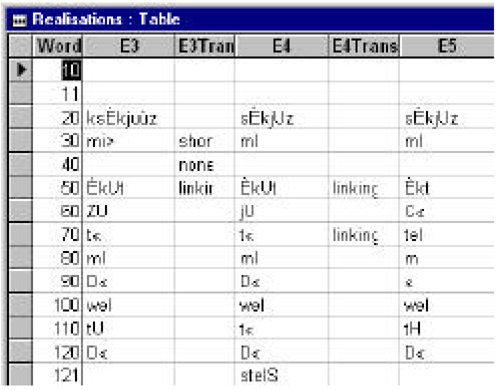
The part that makes a spoken corpus essentially most different from a written one is that not only does a spoken corpus/database need to incorporate the transcriptions themselves, but also needs to give the user access to the original recordings in form of sound files, so that the original transcriptions can easily be verified and potentially corrected. This is why simple annotated text files generally do not represent a suitable storage format for spoken data. More often than not, genuine spoken corpora therefore tend to take on the form of applications that allow linking in and controlling analysis and playback tools (cf. Deutsch et al., 1998).

The methodology I describe in this paper is based around an MS Access 2000 database application that already allows for some of these features, while some others are yet to be implemented. It is, for example, already possible to start a speech analysis program with a specified soundfile and at a specific offset, in order to do transcriptions and to control this program to some extent from a form via Visual Basic for Applications (VBA).

[2] For more information on different types of annotation, see Leech et al. (1998).

3.2. Transcription issues

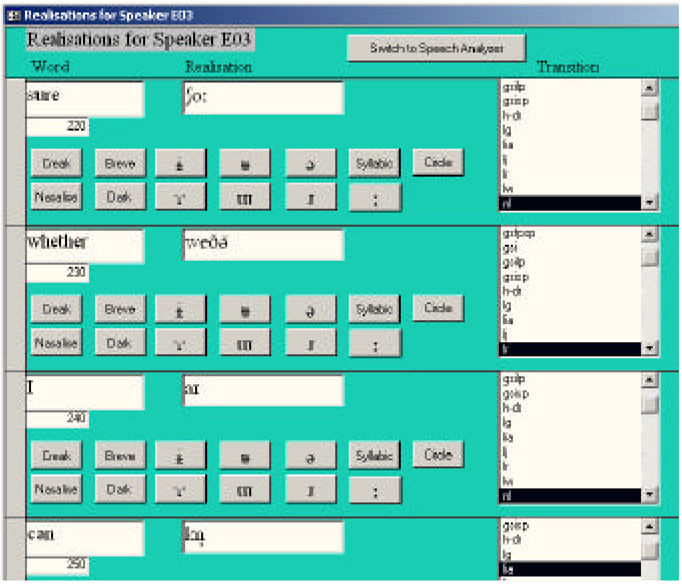
As already mentioned above, the use of fonts may present some difficulties, especially if maximum compatibility between different operating system platforms is required. However, with the increasing importance of Unicode[3] and at least some Linux implementations now supporting the use of True-type fonts, the use of IPA fonts has become less of an issue. Working on a Windows-based system, the obvious choice for my implementation is to use an IPA font whose character representations are saved in the transcription table(s). Since MS Access unfortunately does not allow for multiple fonts within the same table and the font I use only contains a limited amount of numbers, which would therefore make it unsuitable for representing large IDs for linking pronunciations to words, transcriptions and comparisons in the database are made accessible via MS Access forms that can display different fields of a table using different fonts. Figure 1 and Figure 2 below demonstrate the difference between the two forms of representation.



Realisations : Table

Word	E3	E3Tran	E4	E4Trans	E5
10					
11					
20	ksÈkjuùz		sÈkjUz		sÈkjUz
30	mi>	shor	ml		ml
40		none			
50	ÈkUt	linkir	ÈkUt	linking	Èkt
60	ZU		jU		C«
70	t«		1«	linking	tel
80	ml		ml		m
90	D«		D«		«
100	wel		wel		wel
110	tU		1«		1H
120	D«		D«		D«
121			stelS		

Figure 1 – A snapshot of the Realisations table.



Realisations for Speaker E03

Word		Realisation
sure	220	ʃoː
whether	230	ʍeðə
I	240	aɪ
can	250	kn̩

Figure 2 – The Realisations form, open for speaker E03.

As Figure 2 shows, the Realisations form also provides buttons that make it easier to input some of the characters and diacritics that would normally have to be input via character codes as they are not mapped onto the keyboard. Something else that can be seen from the illustration above is that in order to be able to conduct precise phonetic and phonemic analyses, a great amount of detail should be included in the initial transcription. This is necessary because in empirical analyses like mine, and especially when comparing speakers from different countries, it is extremely difficult to predict which

[3] Although existing Unicode fonts still often lack in typographic quality as far as phonetic characters are concerned.

features may turn out to be relevant. For example, after having listened to some of my data, I had initially assumed that one of the major differences between the native and non-native speakers may be the amount of creak in the realisations of the latter, but it later turned out that both speaker populations use creaky voice, while this feature only becomes a distinguishing marker in certain contexts.

A further issue in the transcription spoken data is consistency. In a large scale implementation of my methodology, it would be extremely important to have a sufficient amount of transcriptions verified by different transcribers since this is the only way in which any internal consistency could be guaranteed – and thus also any statistical validity.

3.3. Phonetic categories

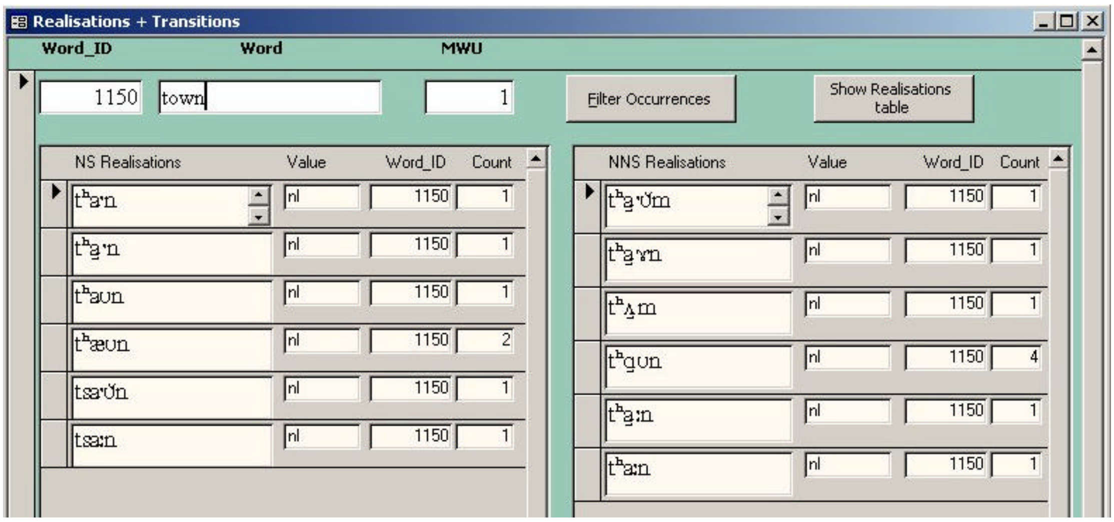
Unlike in a written corpus, where words are generally separated by whitespace or punctuation marks, in transcriptions of spoken data it is important to look at the context in which each word appears. Transitions between words thus represent an important and relatively easily categorisable feature of pronunciation and often serve a distinctive markers between different speaker populations. It is therefore important to keep a record of the phenomena occurring between words alongside the general transcription. Figure 2 above shows a list-box on the right-hand side from which the different transitions can be selected. These transitions contain both simple types, such as assimilation, elision, etc., but also information about pauses and other complex types, such as assimilation + elision, assimilation before a short pause, assimilation before a long pause, etc.

Information about stress patterns can easily be incorporated by including primary and secondary stress marks in the transcriptions, complemented by information about possible multi-word-units, such as compounds, etc. Furthermore, other suprasegmental information, such as the length of the realisation of different objects in the text, i.e. the length of the dialogue (text) itself, of individual sentences, phrases, and down to the level of individual words if necessary. At present, my database(s) only contains information down to the level of individual sentences.

3.4. ‘Phonetics’ vs. ‘phonology’

As previously pointed out, for comparing different speaker populations often a great level of phonetic detail may be required in order to detect relevant distinguishing features. However, this wealth of detail may sometimes make it difficult to arrive at high-level observations. For this reason, my application contains a VBA routine that creates a copy of the original data that can be ‘filtered’ by stripping out any diacritics deemed irrelevant for any particular part of the analysis, in order to make the remaining data more ‘legible’. For example, if it is unimportant for analysis purposes whether initial plosives are aspirated or not, all the aspiration can simply be filtered out in the comparison. In this way, we can move step by step from a phonetic to a more phonemic representation of the data. Figure 3 below shows an ‘unfiltered’ comparison of the realisations and transitions associated with the word *town* and Figure 4 shows a ‘filtered’ representation of the same word with ‘normal’ aspiration removed[4]. The ‘filtering’ routine is started by clicking on the “Filter Occurrences” button at the top of the form displayed in Figure 3, which then prompts the user to input a series of characters to be stripped out, separated by spaces.

[4] Note that the stronger, more unusual aspiration represented by an [σ] is still shown.



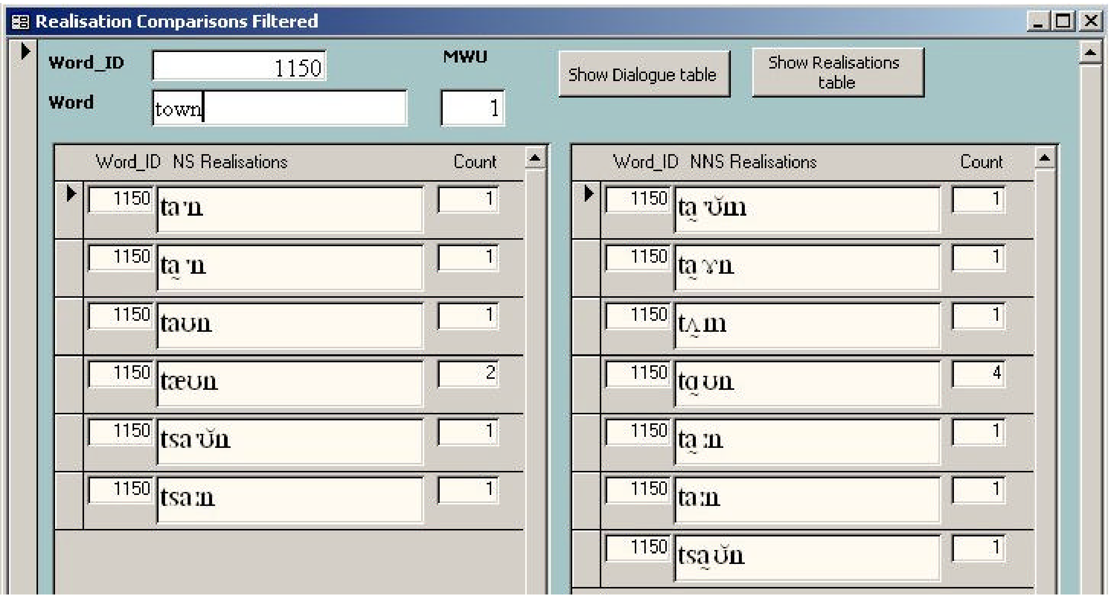
Realisations + Transitions

Word_ID	Word	MWU
1150	town	1

NS Realisations	Value	Word_ID	Count
tʰaˑn	nl	1150	1
tʰa̰ˑn	nl	1150	1
tʰaʊn	nl	1150	1
tʰæʊn	nl	1150	2
tsaˑʊ̆n	nl	1150	1
tsaːn	nl	1150	1

NNS Realisations	Value	Word_ID	Count
tʰa̰ˑʊ̆m	nl	1150	1
tʰa̰ˠn	nl	1150	1
tʰʌ̰m	nl	1150	1
tʰɑ̰ʊn	nl	1150	4
tʰa̰ːn	nl	1150	1
tʰaːn	nl	1150	1

Figure 3 – Realisations + Transitions (‘unfiltered’)



Realisation Comparisons Filtered

Word_ID	1150
Word	town
MWU	1

Word_ID	NS Realisations	Count
1150	ta ˈn	1
1150	ta̰ ˈn	1
1150	taʊn	1
1150	tæʊn	2
1150	tsa ˈʊ̆n	1
1150	tsa ːn	1

Word_ID	NNS Realisations	Count
1150	ta̰ ˈʊ̆m	1
1150	ta̰ ˠn	1
1150	tʌ̰m	1
1150	tɑ̰ ʊn	4
1150	ta̰ ːn	1
1150	ta ːn	1
1150	tsa̰ ʊ̆n	1

Figure 4 – Realisation Comparisons (‘filtered’)

4. Data Storage

Apart from what is being recorded in the kind of corpus under discussion, it is also important to understand how the different types of information are stored and how they can be related to one another in a meaningful way. As far as data storage is concerned, we end up with two fundamentally different types of data, the digitised soundfiles, which are stored as 16 bit, 16kHz .wav files, which can be played back by many player applications both on Windows and Linux systems, and the transcriptions and annotations, which are stored in the relational MS Access database(s).

4.1. Relational model

The choice of using a relational, rather than a flatfile database model was an easy one because a relational database allows to store different types of information together and relate them to each other when needed. It therefore provides a highly flexible and easily expandable architecture to which other types of information can be added if and when necessary.

4.2. Levels of information and their coordination

The most basic levels of information in the database are represented by the transcriptions of individual word tokens, their transitions and the word tokens they are related to in the dialogue. Through this, it is not only possible to combine and display all the different realisations for a particular word in a specific context, but also generate a kind of extended pronouncing dictionary with information about the different possibilities of pronunciation for a specific word in different contexts.

At the next level, we can incorporate grammatical tags in order to see whether specific realisation phenomena, such as for example the deletion of complete words, are related or even restricted to specific word classes for individual speaker populations. We can also investigate whether specific groups of word classes, such as deictica (e.g. pronouns or locatives) are often realised with final release, indicating a potentially unusual degree of emphasis.

Next, we can incorporate punctuation to determine speaker behaviour at phrase, sentence or turn boundaries and see how it relates to the realisation of pauses or final lengthening indicated in the word transition categories. This could then, as a further step, be complemented by prosodic information to investigate the proportion of long or short pauses in relation to the use of prosodic indicators of textual cohesion.

In those cases where we may find particularly unusual realisations that do not fit any speaker population patterns, we can make recourse to the information stored about individual speakers in order to verify whether anything in their backgrounds may have triggered such particular speech behaviour, such as e.g. long periods of time spent abroad, etc. In a similar way, we may also investigate variation within particular populations based upon the age of the speakers.

5. Comparison strategies

Comparisons between different speaker populations can take on various forms. As an initial step, all the realisations (potentially also including the transitions) for each particular population have to be combined into one table. This can be achieved by running SQL union queries against the database. Once the query results exist, additional queries can be run against them and aggregate functions, such as counts, averages or standard deviations can be used to establish frequency information for particular realisations, transitions, etc. Just like with concordance programs for analysing written corpora, wildcard searches in these queries help to focus in on particular parts of the data.

If all that is needed for an analysis is to establish the deviances of one population from another, one can even write programs that analyse the results of the union queries and write these out to a table. However, according to my experience this may not necessarily help the researcher to understand all the relevant differences between the populations. Therefore, contextualised side-by side comparisons, such as illustrated in Figures 3 and 4 above are a much more preferable way of analysing the data, since this way, tendencies for both populations can relatively easily be spotted and compared. Once a particular model for a native speakers population has been established, it is of course also possible to evaluate the performance of individual non-native speakers against this model and to identify in which areas there is scope for improvement.

6. Applications and implications

The methodology I have developed for my PhD thesis provides a way of using relational databases in order to store transcriptions of native and non-native speaker data and for capturing the differences between them statistically. This is a rather different approach from the one taken by most other studies investigating the performance of non-native speakers, who simply assume that a convenient basis for comparison exists in form of a standard.

This kind of methodology has implications for both language teaching and also testing, as it may provide a more realistic account of the way that native speakers speak and therefore also what should potentially be taught to foreign learners in order to enable them to communicate efficiently with the former. As far as language testing is concerned, the implications are that

a) it is possible to assess the speech of learners in a relatively objective way, rather than having to relay on the largely impressionistic marking schemes that are still currently in use (cf. Heaton, 1975: 100 and Weir, 1993: 43/44), and

b) that this provides an eminently better and fairer way of assessing the speech of foreign learners in comparison to native speaker performance.

Apart from its use for the evaluation of non-native speaker accents, the methodology can also easily be applied to the study of different native speaker accents, not only for purely linguistic research purposes, but also potentially in order to establish criteria that may be used to improve speech recognition and other language engineering technology, such as dialogue systems (cf. Leech and Weisser, 2001).

References

Deutsch W, Vollman R, Noll A, Moosmüller S 1998 An Open Systems Approach for an Acoustic-Phonetic Continuous Speech Database: The S_Tools Database-Management System (STDBMS). In: Nerbonne J (ed.) 1998. *Linguistic Databases*. Center for the Study of Language and Information: Stanford, California. pp 77-92.

Chollet G, Cochard J-L, Constantinescu A, Jaboulet C, Langlais P 1998 Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-speaker Variability. In: Nerbonne J (ed.) 1998. *Linguistic Databases*. Center for the Study of Language and Information: Stanford, California. pp 117-135.

Granger S (ed.) 1998 *Learner English on Computer*. London: Longman.

Heaton JB 1975 *Writing English Language Tests*. London: Longman.

Hughes A., Trudgill P 1987 *English Accents and Dialects*. London Edward Arnold.

Leech G, Weisser M forthcoming 2001 Pragmatics and Dialogue in: Mitkov R (ed.). *The Oxford Handbook of Computational Linguistics*. Oxford: OUP.

Leech G, Weisser M, Wilson A, Grice M. 1998 Survey and Guidelines for the Representation and Annotation of Dialogue. In: Gibbon D, Mertins I, Moore R (eds.) 2000 *Handbook of Multimodal and Spoken Dialogue Systems*. Dordrecht: Kluwer Academic Publishers.

Nerbonne J (ed) 1998 *Linguistic Databases*. Center for the Study of Language and Information: Stanford, California.

Wieden W, Nemser W 1991a *The Pronunciation of English in Austria*. Tübingen: Gunter Narr.

Wieden W, Nemser W 1991b Compiling a database on regional features in Austrian-German English. In: Wieden W, Nemser W 1991 *The Pronunciation of English in Austria*. Tübingen: Gunter Narr. pp. 350-363.

Wells, J. 1982. *Accents of English*. Cambridge: CUP. (Vol. 2)

Wells J, Barry W, Grice M, Fourcin A, Gibbon D 1992 Standard Computer-compatible transcription. Esprit project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL.

Weir C 1993 *Understanding and Developing Language tests*. Hemel Hempstead: Prentice Hall International (UK).