

Jacques Durand
Monika Pukli

How to construct a phonological corpus: PRAAT and the PAC project

Résumé

Dans cet article, nous considérons le traitement des données dans un projet moderne comme PAC (Phonologie de l'Anglais Contemporain : usages, variétés et structure). La numérisation des enregistrements, la transcription orthographique du signal, son codage et son analyse sont facilités par les avancées techniques récentes. Nous décrivons ici un outil informatique puissant et cependant convivial appelé PRAAT, qui a été développé par Paul Boersma et David Weenink à l'Université d'Amsterdam. Nous décrivons nos principes de transcription et leur intégration dans cet outil comme phase préliminaire mais indispensable d'une bonne analyse phonologique ou phonétique.

* Cet article est la version préfinale d'un article paru dans le no. 36 de la *Tribune internationale des langues vivantes*, intitulé *English Pronunciation : Accents and Variation*, sous la direction de Philip Carr, Jacques Durand & Monika Pukli. Pour toute citation, se référer à l'article sous sa forme imprimée dans la *Tribune internationale des langues vivantes*, vol. 36 : 36-46.

0. Introduction

Most specialists agree that modern dialectology started with the empirical surveys focusing on urban speakers conducted by Labov in the United States. Going under several labels varying from sociolinguistics and sociophonetics, to urban dialectology and variationist studies, this field of descriptive linguistics has fruitfully incorporated some basic principles from the social sciences and looks at groups of speakers carefully selected along various dimensions (gender, class, age, ethnicity, etc.). Although it is often focused on phonology, it has extended our understanding of other areas such as syntax or discourse structure. This work has gone beyond traditional dialectology which tended to restrict itself to lexical issues (including pronunciation) and was solely focused on rural areas and so-called NORMS (nonmobile, older, rural males: see Chambers & Trudgill 1980: 33-35).

Modern dialectology is also more ‘modern’ from a more practical point of view in the sense that technological development has changed and considerably eased methods and analyses that fieldwork and consequent study of empirical data imply. More specifically, in the domain of phonetic investigation, acoustic studies are much more widely available and many phonologists firmly believe that theory and reliable description can benefit extensively from phonetics and instrumental studies.

Increasing interest in the combination of phonetic and phonological sciences has led to a series of current studies on accent variation and sound change based on empirical investigation, as for instance the collection of research in *Urban Voices* (Foulkes & Docherty 1999), which represents a conscious effort to draw on data and their instrumental examination in order to provide descriptions of various dialects and accents of the British Isles.

The *PAC* project (*La Phonologie de l’Anglais Contemporain: usages, variétés et structure / The Phonology of Contemporary English: usage, varieties and structure*) with its series of joint corpora represents this line of linguistic research of variationist interest uniting phonetic and phonological investigation in the quest to attain an effective and factual description and comparison of various accents of English. Based on a common protocol and following a uniform methodology, the project investigates contemporary English accents in order to 1) give a better picture of spoken English in its unity and diversity (geographical, social and stylistic), 2) test phonological and phonetic models from a synchronic and diachronic point of view, 3) favour communication between speech specialists and phonologists, and 4) provide data and analyses which will help improve the teaching of English as a foreign language. (For a presentation of *PAC* and its methodology see Carr, Durand & Pukli, this volume.)

This paper takes a look at the construction of corpora within the framework of the *PAC* project, and describes what happens in between the moment when oral data is recorded following the standard *PAC* protocol, and the moment when data is made ready for analysis. In Section 1 the different phases of the data collection are described, this is followed by an introductory overview of PRAAT, the principal analytical tool used in the *PAC* project. Section 3 and 4

discuss various principles of data transcription, and the last part is devoted to possible codings of the discourse and software tools employed in the project (Section 5).

1. From data collection to linguistic analysis

Thanks to modern equipment, fieldworking has become much more practical and simple, yielding, at the same time, high-quality data. In *PAC* the use of small DAT recorders (handy and unobtrusive machines) with small clip-on microphones (inconspicuous but very powerful and sensitive devices) has allowed us to obtain digital recordings of a very good quality.

Although working in a studio would ensure superior recording quality in a soundproof setting without any background noise, it would also be less convenient for the informants who would have to come for the recording session (and who are not paid), not to mention the practical inconvenience of having to equip or hire a studio in each location for the survey. And, perhaps most importantly, a studio is more likely to intimidate speakers, and would be certainly less suited to create the four different stylistic settings required for the *PAC* protocol, which includes the reading aloud of two wordlists and a continuous text, as well as conversational speech in formal and casual contexts.

Consequently, using a DAT recorder in the informant's house (or that of a friend of the informant) seems to be a good compromise. Speakers are relatively at ease, conditions are more flexible and comfortable for them, and quality is acceptable. Naturally, everything is done in order to reduce background noise (and one has to be wary of television, radio, electrical appliances, pets, etc.).

The next step involves transferring data from digital recordings to a computer, where it is cut and edited according to certain norms set down in the project to be eventually submitted to the linguistic analysis proper. (For more on fieldwork and corpus construction see e.g. Delais-Roussarie, 2003a,b, and for the French *PFC* project, Durand & Lyche 2003.)

2. Data analysis – PRAAT

Technological evolution of computers and speech analysis software has made handling and analysing data a lot easier. Computers can be used effectively to store and study linguistic data and acoustic analysis has also become accessible to a wider layer of researchers.

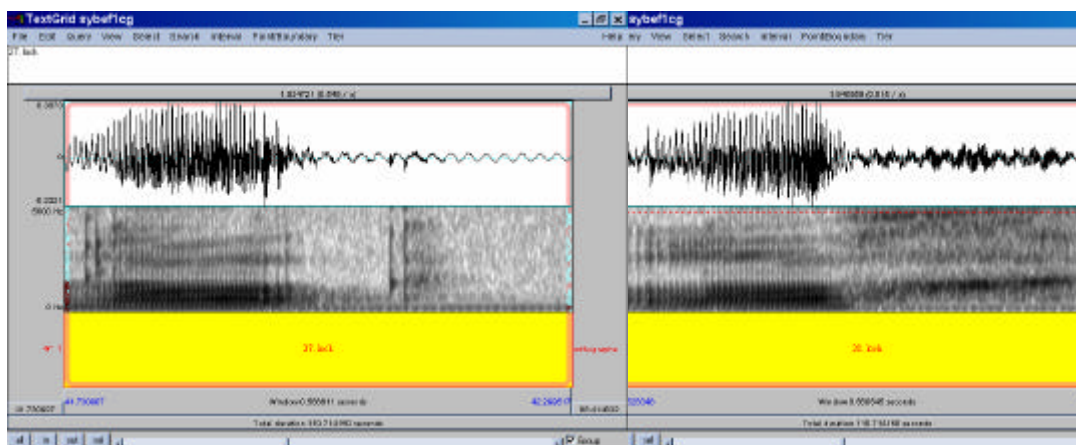
Once recordings are assembled and edited, the corpus is ready to be investigated and the actual data analysis can begin. One way to set about dealing with the extensive and varied input provided by our protocol is to create text files with transcription, commentaries and codings related to various linguistic phenomena in the oral database. The most efficient and practical method we have come across so far is the speech analysis software called PRAAT (the Dutch word for 'speak') developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences at the University of Amsterdam. It is a regularly improved and updated, highly sophisticated tool for speech analysis with manifold functions in a

simple, user-friendly layout. It is also cost-effective, being a shareware program which is downloadable for free use from the PRAAT homepage. It should, however, be mentioned that other annotation and corpus tools could also be appropriate, as for instance the *AGTK: Annotation Graph Toolkit* (from the Linguistic Data Consortium at the University of Pennsylvania), or the *CLAN (Computerized Language Analysis)* program from the Carnegie Mellon University, depending on the main focus of the data collection and the linguistic study (for a review and discussion of these and other tools for corpus construction and analysis see Delais-Roussarie, 2003b.)

With PRAAT we can directly align the audio signal to a text file in a very simple manner, and this in turn permits a convenient manoeuvring of linguistic data: spontaneous speech from a long conversation for instance, can be segmented, labelled and transcribed while the sound itself is displayed (and is available for listening) in the same window as the text files.

We can also have instant access to acoustic analyses in the same working space, and perform a more careful study based on spectrographic analysis with intensity, formant and pitch contours, if necessary. This can prove to be useful for example, when specific information or confirmation is needed for the segmental make-up of an utterance.

To take a fairly simple example, some basic information can be instantly read in picture 1. below of *lock* and *loch*, pronounced in a Scottish accent (see Durand, this vol., Pukli, this vol.). Formant frequencies tell us about vowel articulation: here the closely situated first two formants (dense black horizontal lines) indicate that both words have a low back vowel: [ɔ], while the difference in the nature and duration of the closure in the consonantal sounds shows that *lock* is realised with a short and complete closure (a vertical bar) preceded by ‘silence’ (light grey area between the vowel and the plosive) attributable to a glottal stop: [ʔk], and *loch* has a longer, partial closure (the darker grey area of the noise of friction which is present up until the highest frequencies) : the fricative [x].



Picture 1. *Lock* and *loch* in a Scottish accent: a minimal pair in which the final consonant is realised with a clearly visible

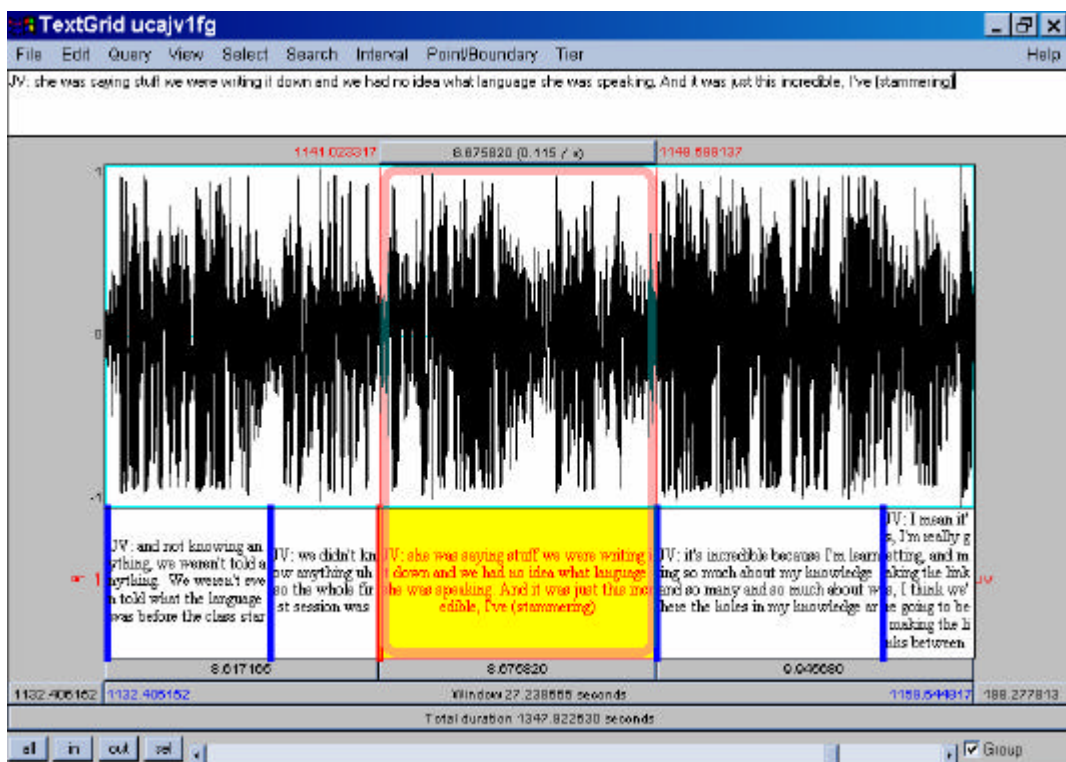
voiceless velar closure preceded by a glottal stop for *lock*, and with a voiceless fricative sound for *loch*.

Further advantages of working with PRAAT include an automatic search function in the text and the audio file (for example, if one is looking for velar nasal realisations in gerunds, or wants to find all occurrences of pronouns for a comparative analysis of strong and weak forms, or is interested in specific dialectal expressions, etc.); and extracting shorter sound sequences from the audio file (if we wished to focus, for instance, on /u/ and its possible phonetic variants in a Scottish accent, we can extract such target words and then compare them across speakers to see whether or not /u/ is subject to variation within that group of informants, or we can examine /u/ and its variants across different corpora to look at regional variation).

Apart from quite simple but useful manoeuvres as those mentioned above, more advanced tasks in acoustic and statistical analyses can also be performed, but they lie outside the scope of the present paper. For further information and discussion on PRAAT (statistics, graphics, speech manipulation, speech synthesis, OT (Optimality Theory) and PRAAT, etc.) see the homepage and its manuals at <http://www.fon.hum.uva.nl/praat/>.

3. Transcription

To complement and facilitate data analysis the input is usually annotated and aligned to readable symbols, i.e. some sort of transcription. This is another and very important phase of an empirical project, which renders the extensive and varied data from spontaneous speech, for example, accessible to linguistic analysis. PRAAT's excellent environment makes transcriptions easy to follow and to explore. At a further stage parallel codings and commentaries are usually added to conversation scripts. Codings can be established to focus on the two main domains of phonological interest within the *PAC* project: T/D and R related phenomena, such as glottaling, tapping, linking, etc. (see Section 5).



Picture 2. PRAAT facilitates data transcription and provides text files aligned with the corresponding sequence of the recording. Further text files, or tiers, are also available to accompany the transcription with comments or codings, or to re-segment the same linguistic input according to other considerations.

In the *PAC* project we have chosen to establish a ‘level zero’ of orthographic transcription, which, for various reasons enumerated below, is preferable to phonetic transcription. This policy has proved profitable in the French *PFC* project (see Durand & Lyche 2003), an experience that prompted the adoption of a similar approach.

Firstly, orthographic transcription helps the researcher assess quickly the linguistic content of the sequences transcribed in the corpus. PRAAT can find occurrences of the specific phenomenon one is interested in, and search based on morphological information (*ing*, *-ed*, *-er*, etc.) can also be effected. Phonetic transcription, on the other hand, will in this sense often ‘destroy’ data. To take a French example, /i/ in a conversation can both stand for *y* and *il*, and the untangling of such and similar neutralizations can hinder quite unnecessarily the analysis of the data.

Secondly, and more importantly, orthography remains neutral as to the phonemic or phonetic make-up of a given variety of language. Transcribing without presuppositions about the phonemic inventory of a given variety is especially useful, if not crucial, since one of the goals of the analysis is precisely to establish phonemic and allophonic distributions for the accent under study. (For a detailed discussion on phonetic vs. phonemic or broad vs. narrow transcriptions, and their complexity, see Durand 2001.)

Thirdly, phonetic transcription is time-consuming; clearly, it is unnecessary for the transcriber to spend ‘hours’ deciding the actual quality of various sounds, in a phase where the specific range of problems to be tackled have not yet been fully defined and with such an important quantity of raw data. As a first step, orthography is more straightforward.

Finally, phonetic transcription implies a fair degree of subjectivity. It represents a form of analysis in which the transcriber’s personal judgement is inevitably involved and the potential inaccuracy of which can easily impede further research. The actual value of IPA symbols and diacritics can also show considerable variation from one transcriber to another.

For large-scale projects orthographic transcriptions have, in fact, become the norm in the initial stage of work on the raw acoustic data. (For more on this see Delais-Roussarie, 2003a, Durand & Lyche 2003.) And later on, if necessary, it can still be followed up by phonemic or phonetic transcriptions, or other codings of the discourse.

Having reliable transcriptions is crucial to subsequent linguistic analysis. This is one of the foremost priorities in *PAC*, therefore thorough and clear conventions were established at the outset of the project in order to guarantee a uniform method of transcription for all corpora. For a sample of the most important principles see the next section.

4. The general orthographic conventions in PAC

In the following some samples are presented taken from Durand & Pukli (2004) to illustrate the conventions for the orthographic transcription followed in the *PAC* project. It is largely based on the principles proposed for the French *PFC* project in Delais-Roussarie et al. (2002); see also Durand & Lyche (2003).

We transcribe under *PRAAT*, on a single tier, regardless of the number of speakers involved. Interval boundaries are added according to the logic of turn taking (a new interval for each new turn). However, if stretches within boundaries are too long, a true phonemic/phonetic alignment may prove difficult at a later stage. Therefore, interval units do not normally exceed 15 seconds. No carriage returns are used, and the speaker is identified at the beginning of each interval.

Simplified punctuation

The punctuation system is simplified: the full stop, the comma and the question mark are the only symbols from traditional spelling used for the transcription of discourse in the project.

JV: I don't know what to do with it, I mean I've never looked at a language that way, which is sort of going out and not knowing anything.

Commas indicate a brief pause in the discourse, or a 'non-final', 'continuing' intonation contour marked by a shift in pitch or other cues.

TB: So I was home. I won the airline tickets.

Full stops stand for a relatively long pause in the discourse, or for a 'final' intonation contour.

DH: How many of these are you going to have?

A question mark is inserted at the end of a question.

NB:

- Pauses and intonation contours do not always coincide with expectations based on syntax.
- Pauses and intonation units are not distinguished along rigorous lines in the orthography employed here; such a finer supra-segmental transcription remains an optional subsequent task.
- Commas are used between repeated words or expressions (cf. 3.1.5).
- An exceptionally long pause in an otherwise logically/syntactically coherent sequence will be indicated by a parenthetical remark (cf. 3.1.6).

LC: but overall I'd say, (silence) a little less than half, of those who apply.

Turn taking

At the beginning of each turn the speaker is identified by his/her initials, which are followed by a colon (a space is inserted on its right, but none on its left). The fieldworker is designated by the letter **F**.

*F: So, do your parents agree with you?
JF: Well, not really.*

As mentioned above, there is no carriage return to mark the end of a sentence or paragraph. The discourse of a single speaker is transcribed continuously under *PRAAT* (with regularly added interval boundaries, each unit being headed by the initials of the speaker).

Overlapping turns

Turns often overlap in a conversation; three types of interventions are distinguished in the transcription:

- Very brief background responses, typical fillers such as ‘yeah’ ‘really’, laughter, vocal and other noises uttered by the listener to maintain interaction are ignored.
- Short interventions – i.e. when the listener interrupts the speaker but does not initiate a new turn, and the speaker goes on speaking – are transcribed within angled brackets in the following manner:

LC: So it's, it's that the approach <F: The approach.> is different.

DR: I mean he may get uh, <F: But Nixon came back. I think if I remember he was beaten once and then.> yeah, yeah that's pretty unusual, pretty unusual.

F: So it's really your grandparents who are Japanese speakers? <JF: Yeah.> Your mum and dad are really English speakers <JF: Yeah.> their, their first language is English?

NB: Here ‘yeah’ has a real ‘response’ value, and, therefore, is transcribed.

- When a listener interrupts the speaker and then ‘takes over’ the conversation, his/her words uttered at the same time as those of the previous speaker are transcribed between angled brackets as indicated above, and a new turn is marked by a new interval (under *PRAAT*).

*F: Do you feel American above all or what do you feel? <TS: Sure I,>
TS: I guess I don't know what that really means, (laughter) I've, you know, I'm an American but, I don't, I'm not like, ‘yeah I'm an American’, you know?*

Truncation of words

A slash (followed by a space) indicates unfinished words:

TS: You think you have this demo/ democratic freedom but it's, not really there.

DH: Well they used to, what ab/ what ab/ what about those uh, the uh, the Concord jets.

LC: the col/ the faculty are looking for a good fit.

JF: My lo/ (laughter) uh it's like non-existent.

Truncated intonation units (when speakers do not finish their train of thought, are interrupted, or hesitate, etc.) are marked by a comma or a full stop:

TS: you know, I am an American but, I don't, I'm not like, 'yeah I'm an American', you know?

Repetition

Repeated words or expressions are separated by a comma.

DR: I, I like to go skiing in the snow, but I don't want to have to dig my way out of it every day.

JF: I think it's true that, that, there is racism in, racism in, in California but it's really well-hidden.

NB: Commas mark repetition and short pauses in the discourse. Thus in the following example the first comma stands for a short pause, the second for a repetition, the third indicates a repetition that coincides with a short pause at the same time, and the fourth one marks a short pause:

JF: Uh, it's okay it's you know it's, it's really, it's really weird teaching you know, I don't know.

Parentheses

Observations made by the transcriber on non-linguistic aspects of the interaction (noises, stammering, laughter, etc.) and on the recording (background action, quality problems) are placed between parentheses.

DG: That's at the beginning of the week so it's hard to remember. (laughter) Uh, we read a couple of theoretical texts comparing irony to allegory,

TB: My father, he is from Canada. (door opens, F returns) Actually he was born in Massachusetts.

Unintelligible words are indicated by the capital letter X in parentheses. The number of Xs inserted (ideally) corresponds to the number of incomprehensible syllables:

JV: because not (XX) all the cases are uh, show up in the pronoun system,

Words are often hard to decipher due to noise or other interferences, in this case the commentaries are inserted in separate parentheses:

RF: kicked everyone out of the airport and made to go you know (noise) (X) shoot the bag and see if it blows up, and uh,

In cases where the transcriber thinks s/he has probably recognized a word (or sequence of words) but is not fully sure, the word is put in parentheses:

JG: Maybe I'll stay in the technology sector, and uh hopefully do something with creativity, like maybe product design, or writing you know (maybe) marketing oriented, something like that. (laughter)

Reported speech

Reported speech is transcribed between inverted commas (‘ ’):

DR: And then when Bush said ‘read my lips no new taxes’ and then, you know,

TB: And there was a woman at the other line and she said, ‘oh no message’, and so I was

TB: and she said I had won the prize and I said ‘didn’t you just call’

Some features of spoken English in relation to spelling

Obviously, many reductions and contractions occur in spontaneous speech. Contracted forms are used in our transcriptions only in so far as they are allowed in standard spelling. Note, in the following example, the co-occurrence of a non-contracted and a contracted form, the former bearing a slight emphasis.

JG: Yeah I have heard that and also I've heard that he seems to be very needy of getting votes.

Sometimes non-contracted forms appear in a more formal style:

F: And were your parents from there?

TS: My mom has lived in Los Angeles all her life.

Word internal ellipsis is an equally frequent feature of spoken English. To avoid a waste of energy at the initial stage of transcription, such deletions are not transcribed. The examination of these features is left to the phonological/phonetic stage of the analysis.

LC: Some very, very intelligent young people, will apply but not do well here because they needed more structure. (and not 'cause)

LC: a portfolio for music, you know original music compositions (and not 'riginal)

But note that we do not reintroduce words (or word sequences) which appear to have been missed out (in relation to normative grammar). Thus if what we hear is:

F: Was she there?

LC: Think so.

We do not transcribe:

F: Was she there?

LC: I think so.

Realizations for which standard orthography offers distinctions will be transcribed accordingly. Thus the distinction between *yes* and *yeah* is systematically respected in the transcription.

TS: I don't know. Yeah. <F: It's confusing.> It's confusing. (laughter) Yes there, there's a lot involved and I think, to be, to say a real opinion on it you have, I have to be, really informed.

F: But do you feel now you're from California ? <TS: Yes.> That you're Californian? <TS: Yes. I guess. (laughter)>

Interjections are another characteristic feature of conversations, employed to express pain, surprise (*ouch, oops*) etc., or simply to provide feedback and to signal active participation towards the other party in the discourse (*uh huh, oh, ah, hm*). For these speech forms, we use the conventions put forward in the OED.

Most often, however, the speaker is simply using a filler to gain time while thinking, hesitating, or searching for an expression (*hm, uhm, uh, er*) etc. Regardless of the actual sound pronounced, this type of intervention will always be described as 'er' for British, and 'uh' for American speech.

RM: Er, it's, it's, er, yeah, it was quite a nice place er, (XX) smelly in some places, the (XX) particularly, er it's very run down and er

DG: Uh, let's see, uh, I uh, I'm from L.A. and I let's say I've been moved uh always to magnet schools which are like schools that kind of specialize in one thing or another

Acronyms

Acronyms – pronounceable words made up from the initial letters of a multi-word name like, for example, UNESCO for the United Nations Educational, Scientific and Cultural Organization – are written in the usual way: capital letters with no separation of any kind if the word is pronounced as a unit. If on the other hand it is spelled out letter by letter, this is indicated by writing a full stop after each letter of the word: U.N.E.S.C.O.

Any unexpected form of actual pronunciation will be indicated in parentheses after the word in SAMPA transcription. SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet developed by speech researchers from many different countries in the late eighties (see www.phon.ucl.ac.uk/home/sampa). It is to date the best international collaborative basis for a standard machine-readable encoding of phonetic notation mapping the symbols of the International Phonetic Alphabet onto ASCII codes. As with the ordinary IPA, a string of SAMPA symbols does not require spaces between successive symbols. (For an example of SAMPA transcription cf. next section.)

Dialectal expressions

Words or expressions that do not belong to either standard British or American English will be transcribed by using SAMPA symbols.

LC: If you want to tell someone to shut up, you say '(hod j@r wiSt)'.

However, if there is a longer stretch of discourse in dialectal speech, "normal" spelling will be employed. If there is a reference dictionary of the dialect being described, its conventions should be used.

Reference orthographic systems

In our transcriptions, we apply the spelling system normally used in the country where the speakers live or come from. Thus, if we transcribe British varieties of English, we use standard British English conventions (adopted in the OED). If we transcribe American English, we use the conventions adopted in Webster's (cf. hesitation 3.1.8). Examples transcribed according to the British and the American conventions, respectively:

Standard British English:

DR: he can't honour the guidelines of the debate for even ninety minutes

RF: when I was, I think, maybe thirteen, just travelling with my mum, and er

American English:

DR: he can't honor the guidelines of the debate for even ninety minutes

RF: when I was, I think, maybe thirteen, just traveling with my mom, and uh

5. Codings and other tools

As mentioned in Carr, Durand & Pukli (this vol.), once orthographical transcriptions aligned with the acoustic signal are available various techniques can be applied to describe and analyse the data. In addition to the techniques listed there and the tools devised within the PFC project, it might be worth mentioning:

TransPraat

TransPraat, a program created by A. Meqqori at the ERSS-CNRS (University of Toulouse 2), converts the coded text files created by PRAAT into 'normal' documents. Naturally, from that moment on it is no longer linked to the original audio signal or its source file under PRAAT, and the transcriptions appear as a continuous text without any codes, figures and symbols.

Variation analysis

Automatic analysis of variation across speakers will also be possible with the help of a software developed for wordlists. The '*Comparer*' can handle up to one hundred words per informant, and items can be listened to and compared from

one speaker to the other. This facilitates impressionistic evaluation of the words from the lists, and most importantly, one can contrast selected speakers on the basis of gender, age and regional criteria (e.g. a comparison of 20+, male and female speakers, or of 60+ speakers from Ayr and Edinburgh).

Akustyk for PRAAT

Akustyk for PRAAT is a program developed by Bartłomiej Plichta at the Michigan State University (cf. <http://bartus.org/akustyk/>). It is an advanced tool for the acoustic analysis of vowel sounds, to be installed and used within PRAAT. Among other things, it provides automatic formant plotting and analysis in a particularly well-organised framework of project management.

6. Conclusion

In this article, we have focused on the treatment and transcription of data within the *PAC* project as an example of the possibilities offered by advances in NLP (Natural Language Processing). Naturally, what is crucial for the linguist is the description and theoretical interpretation of the data. Our simple claim is that there are no good theories without good data. The explicit procedures followed within the *PAC* project do not guarantee success in either domain but seem to us an essential step towards better theoretical accounts of the phonology of even well-described languages such as English and French.

Jacques Durand

CNRS-UMR 5610, ERSS, Université de Toulouse-Le Mirail

Monika Pukli

CNRS-UMR 5610, ERSS, Université de Toulouse-Le Mirail

References

- AGTK: Annotation Graph Toolkit*. Online. <http://agtk.sourceforge.net/>
Boersma, Paul & David Weenink *PRAAT: doing Phonetics by computer*.
Online. <http://www.fon.hum.uva.nl/praat/>
Carr, Philip, Jacques Durand & Monika Pukli (this volume). *The PAC project: principles and methods*.
CLAN (Computerized Language Analysis). Online.
<http://childes.psy.cmu.edu/>

Delais-Roussarie, Elisabeth (2003a). Constitution et annotation de corpus : méthodes et recommandations. In E. Delais-Roussarie & J. Durand (eds.) *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Universitaires du Mirail. 89-126.

Delais-Roussarie, Elisabeth (2003b). Quelques outils d'aide à la transcription et à l'annotation de données audio pour constituer des corpus oraux. In E. Delais-Roussarie & J. Durand (eds.) *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Universitaires du Mirail. 127-158.

Delais-Roussarie, Elisabeth, Jacques Durand, Chantal Lyche, Abderrahim Meqqori et Jean-Michel Tarrier (2002). Transcriptions des données : outils et conventions. *Bulletin PFC 1* : 21-34. CNRS ERSS-UMR5610 et Université de Toulouse Le Mirail.

Delais-Roussarie, Elisabeth & Jacques Durand (eds.) (2003). *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Universitaires du Mirail.

Delais-Roussarie, Elisabeth, Abderrahim Meqqori & Jean-Michel Tarrier (2003). Annoter et segmenter des données de parole sous PRAAT. In E. Delais-Roussarie & J. Durand (eds.) *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Universitaires du Mirail.

Durand, Jacques (2001). La transcription phonétique de l'anglais : problèmes théoriques et pratiques', in W. Rotgé (ed.) *Agrégation externe d'anglais : mode d'emploi*. Paris: Ellipses, 181-219.

Durand, Jacques (this volume). English in early 21st century Scotland: a phonological perspective.

Durand, Jacques & Chantal Lyche (2003). Le projet 'Phonologie du Français Contemporain' (PFC) et sa méthodologie. In E. Delais-Roussarie & J. Durand (eds.) *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Universitaires du Mirail.

Durand, Jacques & Monika Pukli (2004). PAC project: methods and tools. *Carnets de grammaire*. CNRS et Université de Toulouse-Le Mirail: Rapports internes de l'ERSS, vol. 14.

Foulkes, Paul & Gerard Docherty (eds.) (1999). *Urban voices. Variation and change in British accents*. London: Edward Arnold.

La Phonologie du Français Contemporain. Online. <http://infolang.u-paris10.fr/pfc/>

Meqqori, Abderrahim & Jacques Durand (2003). Manuel d'utilisation de Transpraat. *Bulletin PFC 2* : 39. CNRS ERSS-UMR5610 et Université de Toulouse-Le Mirail.

Plichta, Bartłomiej *Akustyk for PRAAT*. <http://bartus.org/akustyk/>

Pukli, Monika (this volume). *Scottish English and the Scottish Vowel Length Rule – An empirical study of Ayrshire speakers*.

SAMPA – computer readable phonetic alphabet. Online. www.phon.ucl.ac.uk/home/sampa