

SEGRE: AN AUTOMATIC TOOL FOR GRAPHEME-TO-ALLOPHONE TRANSCRIPTION IN CATALAN

P. Pachès¹, C. de la Mota², M. Riera², M. P. Perea³, A. Febrer¹
M. Estruch², J. M. Garrido², M. J. Machuca², A. Ríos², J. Llisterrí²,
I. Esquerra¹, J. Hernando¹, J. Padrell¹, C. Nadeu¹

Affiliations

- (1) Universitat Politècnica de Catalunya
- (2) Universitat Autònoma de Barcelona
- (3) Universitat de Barcelona

E-mail

- (1) {carne, montse, monica, juanma, maria, mestre, joaquim}@liceu.uab.es
- (2) mpperea@lincat.ub.es
- (3) {paupag, ignasi, febrer, javier, jaume, climent}@gps.tsc.upc.es

Postal

- (1) Departament de Teoria del Senyal i Comunicacions
Campus Nord, D5, Universitat Politècnica de Catalunya
Jordi Girona, 1-3
08034 Barcelona, Spain
- (2) Departament de Filologia Espanyola
Edifici B, Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain
- (3) Departament de Filologia Catalana
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
08071 Barcelona, Spain

Phone

- (1) 34 93 401 6438
- (2) 34 93 581 1912
- (3) 34 93 403 5614

Fax

- (1) 34 93 401 6447
- (2) 34 93 581 1686
- (3) 34 93 403 5698

Pages: 10

SEGRE: AN AUTOMATIC TOOL FOR GRAPHEME-TO-ALLOPHONE TRANSCRIPTION IN CATALAN

P. Pachès¹, C. de la Mota², M. Riera², M. P. Perea³, A. Febrer¹
M. Estruch², J. M. Garrido², M. J. Machuca², A. Ríos², J. Llisterri²,
I. Esquerra¹, J. Hernando¹, J. Padrell¹, C. Nadeu¹

Universitat Politècnica de Catalunya⁽¹⁾, Universitat Autònoma de Barcelona⁽²⁾, Universitat de Barcelona⁽³⁾
E-mail: {paupag, ignasi, febrer, javier, jaume, climent}@gps.tsc.upc.es⁽¹⁾
{carne, montse, monica, juanma, maria, mestre, joaquim}@liceu.uab.es⁽²⁾, pperea@lincat.ub.es⁽³⁾

Abstract

Segre is a rule-based automatic phonetic transcription system for Catalan, jointly developed by the Universitat Politècnica de Catalunya, the Universitat Autònoma de Barcelona and the Universitat de Barcelona in the framework of the Catalan Reference Centre for Language Engineering (CREL, Centre de Referència en Enginyeria Lingüística).

The syntax of the rules has been designed to obtain phonetic transcriptions for four major dialects of Catalan: the Central or Eastern dialect, spoken in the East of Catalonia, the North-Western or Western dialect, spoken in the West of Catalonia (including the South), the Balearic, spoken in the Balearic Islands, and finally the Valencian, spoken in the Valencian Community.

The transcriber has been designed in a very flexible way, since the rules are fed to the program, which has very little hardwired knowledge. They are defined externally and specified in a number of ASCII text files following a simple syntax for grapheme-to-allophone and allophone-to-allophone conversion rules, the latter necessary to obtain those modifications due to coarticulation phenomena across word boundaries. So, the tool provides a phonetic transcription, as broad or narrow as desired, in isolated mode (without coarticulation across word boundaries) or in text mode (with coarticulation between words).

Furthermore, and due to the fact that the rules may be tweaked in any desired way, it is also possible, for instance, to transcribe particular subdialects or to obtain more or less narrow transcriptions. This follows from the fact that there is not a closed list of allophones in terms of which words are transcribed. Instead, the allophones are given by the various rule files.

The accuracy of transcriptions of new texts, when compared with human expert generated transcriptions, is of 99.1% for isolated words and 99.39% for running text.

Segre can be then considered a useful tool to model how coarticulation modifies the isolated transcription of words in real sentences. So, it is helpful not only to build speech synthesis systems but also to train subword-based speech recognition systems. Certainly, although in simple tasks such as connected digits or phonetic recognition no phonetic dictionary is needed, if the coarticulation rules are incorporated into the recognition network, they may complement the work done by cross-word units.

INTRODUCTION

Phonetic transcription is necessary in applications that operate on phonemes or allophones. For instance, the generation of synthetic speech needs the phonetic information that can be derived from the input text. And most speech recognition system architectures are based on a sub-word level and therefore on the phonetic transcription of a text or a vocabulary. In that context of possible applications *Segre* is a useful tool as it obtains automatically a phonetic transcription from an input text.

Segre is a rule-based automatic phonetic transcriber for Catalan, jointly developed by the Universitat Politècnica de Catalunya, the Universitat Autònoma de Barcelona and the Universitat de Barcelona in the framework of the Catalan Reference Centre for Language Engineering (CREL, *Centre de Referència en Enginyeria Lingüística*). As described below, the nature of Catalan language (and its considered dialects) allows the definition of a set of rules (with their exceptions) that specify the mapping from graphemes to allophones. Besides, coarticulation phenomenon can also be described by a set of allophone-to-allophone conversion rules.

However, neither grapheme-to-allophone (GtoA) nor allophone-to-allophone (AtoA) conversion rules have always a unique definition. That is, phonetic transcription can be as narrow or broad as desired, depending on the set of allophones considered. Additionally, transcription rules vary in different dialects or even speaking styles or rates. These facts motivate the design of a system with minimal hardwired knowledge. Conversion rules are fed to the program in a number of ASCII files specifying rules in a simple syntax as well as information about syllable segmentation, prefixes, monosyllable words and general exceptions. Human expert generated rules are easy and immediate to test, as they are external to the code of the program.

Once this knowledge has been loaded, the transcription of each isolated word is done in several steps:

1. The word is segmented into parts of word when it has an apostrophe, a dash or starts with a prefix.
2. Each part of word is segmented into graphemic syllables by finding the nuclei of the syllables and distributing the intervening graphemes between the different syllables.
3. Each word is accented (in Catalan there are fixed rules to do this step).
4. GtoA rules are applied from left to right grapheme by grapheme, resulting in a phonetic transcription of each word part, including stress marks (primary or secondary) and syllable boundaries.
5. AtoA (or coarticulation) rules are applied between boundary allophones of each word part.
6. Syllabic redistribution rules apply to those words having more than one component part.

The *Segre* tool obtains the connected mode phonetic transcription by applying the following procedure to a given input text:

1. Punctuation marks (loaded as an argument of the program) are used to identify sentences and therefore they are treated as silences. Each sentence is processed separately.
2. Each word is transcribed in isolated mode as defined above.

3. Allophones of the word boundaries may suffer from coarticulatory effects, which are defined in the AtoA rules.
4. The application of syllabic redistribution rules is the last step to obtain the final phonetic transcription.

Rule files mentioned above consist of a list of individual rules that are applied one after the other by order of appearance. If an observed rule applies the following rules are ignored and next grapheme (GtoA) or allophone (AtoA) is analyzed. Keywords (to describe groups of allophones) can be defined at the beginning of files to allow rules to be more compact. The syntax of these rules is described as follows:

a. Grapheme-to-allophone rules

<graph> [alloph] rule_predicate

Each rule specifies to which allophone *alloph* corresponds the grapheme *graph* (how it is pronounced) in the conditions given by the predicate of the rule. Rule predicates are a list of subpredicates indicating particular conditions such as stress, word or syllable position, and preceding and/or following graphemes, among others.

b. Allophone-to-allophone (coarticulation) rules

[alloph1] [alloph2] rule_predicate

Similarly to GtoA rules, it indicates which allophone *alloph2* (which can be empty) substitutes the allophone *alloph1* in the conditions given by the predicate of the rule. Predicates are fairly the same.

c. Syllabic redistribution rules

[alloph] [action|direction] rule_predicate

Accounting for syllabic redistribution, *alloph* is the allophone to which the rule may apply in the conditions given by the predicate of the rule. **The action can be *sil* (syllable) or *fon* (allophone) depending on what is moved to the preceding or following syllable, and *direction* can be *esq* (to the left or preceding syllable) or *drt* (to the right or following syllable).**

DESIGNING RULES TO OBTAIN PHONETIC TRANSCRIPTIONS FOR CENTRAL CATALAN

Although Catalan has a fairly straightforward mapping between graphemes and allophones, its spelling system fails to provide an adequate representation of speech, since there is not a regular one-to-one correspondence between orthographic and phonetic units. For instance, the two graphemes <z> and <s> in words like *zero* [z`Eru] ('zero') and *rosa* [r`Oz@] ('rose') stand for one rather than two sounds, while the single grapheme <g> stands for two different sounds in words like *gat* [g`at] ('cat') and *geniva* [Z@n`iB@] ('gum'). If a text is to be

transcribed, the engine chooses the right conversion among the different possible phonetic symbols linked to each grapheme. This is achieved by using both phonological and statistical information.

-First, phonemic rules fill in predictable information about the way in which the allophones are realised in different environments.

-Second, the remaining cases can be described by means of contextual rules and lists of items. This information is obtained from human expert generated transcriptions of the 67000 entries of the dictionary published by the *Institut d'Estudis Catalans*.

The output of the transcriber is in the form of standard SAMPA (*SAM Phonetic Alphabet*) symbols (Wells, 1987, 1989, 1995)

1. Vowels

The vowel system of Central Catalan has seven vowel phonemes (/i/, /e/, /E/, /a/, /O/, /o/, /u/), but there are rules which neutralise vowel distinctions in certain environments, crucially, according to the stress placement. Thus, unstressed vowels /e/, /E/ and /a/ are in most cases realised as [ə], while unstressed /O/, /o/ and /u/ are commonly realised as [u]. The point is that sometimes this rule is not regularly applied.

a. Stressed vowels

Although there are usually no problems in identifying the phonetic realisation of stressed vowels, there are some phenomena that affect the features of mid low vowels.

-Pre-stressed suffixes, for example, can affect stress placement in the base form to which they are attached. They also force the open realisations [E] and [O] (i.e. *direct*[^ˈo]r-*direct* [^ˈO]ri, *núm*[^ˈə]ro-*num*[^ˈE]ric).

-Other cases depend on diachronic processes. The current nature of <e> and <o> depends on the nature of the Latin vowels. We find then alternations, which are arbitrary from a synchronic point of view. The transcriber, since it has no historic information, uses word endings to obtain the right phonetic representation. Some of them are suffixes (*passeig*), but others are not (*corolla*).

-There are also cases that can be explained using phonetic contextual information, like *perla*, *terra*, *borla*, *orfe*; they are transcribed with stressed open vowels since they are followed by /rr/, which typically triggers vowel opening.

-Some remaining contexts used by the rules have been obtained analysing the global tendency of several strings to be realised in a certain way. That is, words beginning with <lle> tend to be pronounced as beginning with an open vowel: <L`E>.

All the cases that do not follow the proposed rules are treated by means of word lists.

b. Unstressed vowels

Unstressed vowels can also show absence of vowel reduction. In certain cases, /E/ and /e/, instead of being realised as [ə], can be realised as [e]. In a similar way, /O/ and /o/ can be pronounced as [o] instead of becoming [u]. However, there are lexical, morphological and contextual restrictions to the vowel reduction rule.

-Lexical constraints. This is the case in loans (*bitter*, *bàsquet*, *judo*, *adàgic*, *al·legro*), toponyms (*Andes*, *Mònaco*, *Oslo*, *Chicago*, *Londres*) or anthroponyms (*Montse*, *Tere*, *Cloe*), among others.

-Morphological constraints. The transcription system avoids vowel reduction of stressed vowels in the first member of compounds, in stressed prefixes and in adverbs ending in *-ment*. Phonological vowel reduction is supposed to apply after deaccenting rules. So, since these pieces are not deaccented, the vowel reduction rule does not affect them (*trencaclosques*, *semicercle*, *primerament*).

-Contextual constraints. For instance, unstressed <e> preceding [a] is necessarily realised as [e], as in *balneari*, *teatre*, *real*. This can be also the solution in words obtained by derivational affixation: *teatralitat*, *realisme*.

The *Segre* tool transcribes [ə] when neutralisation is not rule-governed but speaker-dependent, as the application of the reduction rule to unstressed vowels seems to be the unmarked option in Central Catalan.

Some phonological processes, crucially those related to adjacent vowels, are sensitive to the occurrence of particular types of boundary. Thus, shortenings, deletions and glide formations are often triggered between base forms, between prefixes and bases, and across word boundaries (*antiinflamatori*, *microordinador*, *portaavions*, *poliesportiu*, *canvi important*, *torna aviat*, *primer i últim*, *menjar enciam*,...). The transcriber provides the most common output for each environment. However, when other prosodic, syntactic, categorial, or arbitrary lexical information is needed to predict the result properly, the hiatus is preferred, since usually this is the recommended realisation in formal speech.

2. Consonants

Consonants are usually specified for three features: voicing, place of articulation and manner of articulation. Phonemic representations are mapped onto phonetic representations by means of a set of phonological rules. Their outputs form the phonological derivations. In spite of the effects of these rules upon one another, even more common in processes affecting consonants, the *Segre* tool identifies the final output within a given domain directly. Each rule specifies to which allophone corresponds the grapheme in the stated conditions. Therefore, when there are several rules related to a grapheme, the conversion rules that have higher priority apply sequentially before those with lower priority, i. e., a

more specific rule applies before the more general rule. The rules must be thus carefully ordered so as to ensure that their interaction is correctly handled.

The phonological system of Central Catalan is governed by different rules affecting consonants, including strengthening, assimilation or deletion processes.

a. BL and GL clusters strengthening

Geminate consonants take the highest rank of the phonemic strength hierarchy. The first consonant of the clusters BL and GL is commonly geminated in Central Catalan (Bonet & Lloret, 1998: 93), although the realisation as an approximant or as a geminate stop depends mainly on the stress position and on word formation (*po[bb]le-po[bb]let*, *su[B]lim-su[B]limar*). However, there is a high degree of variation. To account for this alternation both rules and exhaustive lists of words are required.

b. Strengthening: consonant-final affrication

Devoiced affricates in word final position are the phonetic form of underlying affricates (*lle[tS]-lle[dZ]a*) or fricatives (*ra[tS]-ra[Z]ar*). There are two phonological processes involved: affrication and devoicing, and they are sensitive to the effect of clitic boundaries (*fu[Z]i*, *fu[Z]-hi*, *fu[tS]*, *fu[dZ] ara*). Regulation of rule interaction is then required, since one rule can affect the potential input to another rule. This is easier for the automatic transcriber when graphemes are divergent: *pu[Z]ol (pujol)-pu[tS] (puig)*; *fu[Z]i (fugi)-fu[tS] (fuig)*. However, sometimes graphemes are coincident: *pu[tS] (puig)-pu[dZ]mal (Puigmal)*; *fu[tS] (fuig)-fu[Z]-hi (fuig-hi)*.

Using our rule formalism, we may express this as:

```
<g> [tS] -davant <i> -tonica -posmot F -fitxer_exc ../central/excepcions/exc_zig-menys_excepcions (file: cen_regles)
```

```
<g> [tS] -in front of <i> -stressed -word position Final -file with exceptions ../central/exceptions/exc_zig-except the exceptions (file: central_rules)
```

This applies to cases like *puig* or *fuig*. Before clitics, there is no affrication or devoicing at all:

```
<g> [Z] -davant <i> -tonica -seguit <-h> (file: cen_regles)
```

```
<g> [Z] -in front of <i> -stressed -followed by <-h> (file: central_rules)
```

And, finally, the rule:

```
[tS] [dZ] -darrere F_SONOR SEMICONS (files: cenff_intra_regles and cenff_regles)
```

```
[tS] [dZ] -behind F_VOICED GLIDE (files: central AtoA_intra_rules and central AtoA_rules)
```

deals with the voicing assimilation in compounds (*Pu[dZ]mal*) and across word boundaries (*pu[dZ] meravellós*).

c. Final devoicing

Underlying word final obstruents (the class of oral stops, affricates and fricatives) are devoiced in word final position (*tu[p]-tu[B]et*). Forms like *roc* (*roca*, *roqueta*) and *poc* (*poca*, *poquet*) which end in voiceless have the same phonetic form in spite of their different underlying forms. Besides, the final position is also affected by the voicing assimilation (*pe[z]ar*, *pe[s]*, *pe[z]* *desitjat*, *pe[z]* *observat*).

The Segre tool transcribes orthographic symbols, so, it obtains independently *pe[z]* *desitjat* and *pe[s]*, which are in different environments:

pe[z] *desitjat* : *[s]* *[z]* -*darrere F_SONOR SEMICONS F_VOCAL* (file: *cenff_intra_regles*)
 [s] *[z]* -*behind F_VOICED GLIDE F_VOWEL* (file: *central_AtoA_intra_rules*)

pe[s]: <*s*> *[s]* -*defecte* (file: *cen_regles*)
 <*s*> *[s]* -*default* (file: *central_rules*)

Alberg and *albergs*, for instance, are transcribed taking into account the application of final devoicing, but *grocs* (*gro[G]a*) and *pocs* (*po[k]a*) are not considered for the devoicing rule, since the grapheme <*c*> is usually linked to voiceless consonants.

d. Rhotics

Underlying rhotics can be realised as trills, flaps or elisions, according to the context. In general, flaps are followed by vowels or <*b,v,d,g*>, i. e., approximants (*ca[r]gol*, *ca[r]bó*, *pè[r]dua*), and trills are found in syllabic coda position preceding consonants other than approximants and also in word final position of oxiton items before pause (*mar*).

e. Weakening: stop spirantisation

In some cases stops can reduce in degree of stricture by becoming an approximant, so that underlying stops /*b*/, /*d*/ and /*g*/ can become approximants or stops at the phonetic representation. Approximants and stops appear in complementary distribution due to its allophonic nature. The occurrence of one allophone rather than the other is then fully predictable from the context, being a general phenomenon that applies within the word and across different words.

f. Assimilations

Voicing assimilation

Voiceless consonants may acquire some voicing when they are adjacent to inherently voiced segments like vowels or sonorants. In Central Catalan, fricatives and affricates in coda position become voiced before vowels (*pa[z]* *estret* - *pa[s]ar*, *despa[dz]* *obert* - *despa[tS]et*). This rule applies before resyllabification. Moreover, obstruents in coda position agree in voicing with the following consonant. This explains the voiced stop in *po[d]* *netejar* and the voiceless stop in *po[t]* *sortir*, both with an underlying voiced consonant

(*po[D]er*). The *Segre* tool transcribes *po[d]* *netejar* and *po[t]* *sortir* independently, since the rules apply to the same grapheme <t> but in different environments.

Coarticulatory effects are not just limited to the last phone in a word. On the contrary, the whole coda is voiced in cases like *po[gz]* *bolígrafs*, *fe[dz]* *normals*, *to[dz]* *nets*. Since the tool is able to identify syllable boundaries and syllable nuclei, these codas are correctly transcribed with voiced consonants.

Place of articulation assimilation

In Central Catalan there are many assimilations in place of articulation, specially involving alveolar (/n/ and /l/) and dental (/t/ and /d/) consonants in coda position. Coronals tend easily to assimilate to the place of articulation of the following consonant, even across word boundaries (*i[m]ventar*, *vindra[m] pintors*, *po[p]* *pintar*, *fe[k]* *conegut*). This is always the case when assimilations affect nasal consonants, and it is also the only possibility for certain frequent words like *fu[bb]ol* and *se[mm]ana*. However, although all these assimilations involving other coronals are really common in casual speech, they are not supposed to be systematic in formal styles (Bonet & Lloret, 1998: 144; Recasens, 1993: 180). Accordingly, the transcriber avoids triggering assimilations in place of articulation in most cases.

Manner of articulation assimilation

Homorganicity (i. e., the fact of sharing the same place of articulation) is the main requirement to allow manner of articulation assimilations in Central Catalan. The homorganic nasal and lateral assimilation process affecting stops in coda position depends on factors like tempo, speaking style and the syntactic cohesion between the words in which the segments are embedded (Bonet & Lloret, 1998: 149). So, the engine does not apply this rule across word boundaries. Some usual words like *espa[LL]a* and *emmo[LL]ar*, which are supposed to have underlying geminates, are transcribed with gemination.

g. Continuant simplification

Rhotics and sibilants, which are specified as continuant and coronal, can be simplified into a single segment when they are adjacent, partly because of the tendency to avoid sequences of identical phonological objects (*é[Z]aponès*, *mé[S]op*, *de[s]alar*). When a sibilant is followed by a rhotic is then possible to rhotize the sibilant: *l[rr]ael*. However, this is not the preferred solution in formal speech (Bonet & Lloret, 1998: 170). Therefore, this is not the output obtained from the automatic transcriber.

h. Deletions

Underlying /n/ and rhotics in final position undergo the most extreme form of phonological weakening, that is, deletion. When the deleted consonant is /n/, final consonant deletion is not a real problem for the transcriber, since the output of the rule is always represented orthographically (*pa-panet*). Final rhotics, however, are represented by means of orthographic symbols even in the cases subject to elision (*cantar*). Although the environments triggering the process are related to the stress, the rules regulating the pronunciation or omission of rhotics are in fact complex. There are many exceptions and a

significant amount of variation among speakers. For this reason, some cases are transcribed using specific lists.

Elision of stops in word final clusters is systematic when the first consonant is nasal or lateral and the consonants are produced at the same major place of articulation. This is then the case of *ca[m]* (*camperol*), *mo[l]* (*molta*), *fa[N]* (*fangós*). In those environments where the elision is optional, the transcriber maintains the cluster, since this is the preferred output in formal speech.

3. Syllabification

Initial syllabification takes place in the lexicon, since the syllable domain is relevant within the lexical phonology. Besides, there may be a level of postlexical resyllabification, relevant for processes that operate across word boundaries. At first, the Segre system specifies stress marks and syllable boundaries, and a set of redistribution rules applies afterwards to perform further resyllabification. The resyllabification notion implies that sequences of the sort CVCVCV, as in *mar era*, will be syllabified into a sequence of CV syllables, rather than, say, into one CVC syllable (*mar*) followed by a V syllable (*e*) and a CV syllable (*ra*). Syllables with coda preceding syllables without onset are not allowed. In those cases the consonant in coda position is linked to the next syllable.

CONCLUSIONS

The Segre engine is a flexible system, with very little hardwired knowledge and a set of rules fed to the program, which transcribes phonetically peculiar subdialects. Possible improvements include a parsing of the sentences to distinguish some words that have a different pronunciation depending on their grammatical category (verbs, nouns, and so on). This would contribute very little to the correctness of the transcriptions obtained automatically, since the accuracy of automatic transcriptions of new texts, when compared with human expert generated transcriptions, is of 99.1% for isolated words and 99,39% for running text.

Con formato

EXAMPLE

Accepten a ulls clucs els ídols esportius, artístics o aristocràtics que els venen els agents publicitaris.

(Eulàlia Solé, AVUI, 13-01-2000)

@k|s`ep|t@|n@|`uLs|kl`ug|z@|z`i|Du|z@s|purr|t`iuws
@rr|t`is|tig|z`O|ris|tu|kr`a|tik|s|k@|lz|B`E|n@|n@|z@|Z`
ens|pu|Bli|si|t`a|ris

BIBLIOGRAPHY

- BONET, E. & LLORET, M.R. (1998) *Fonologia catalana*, Ariel lingüística, Barcelona.
- CASTELLANOS, J.A. (1993) *Manual de pronunciació*, Eumo Editorial, Vic.
- IEC (1999) *Proposta per a un estàndard oral de la llengua catalana, I, Fonètica*, Institut d'Estudis Catalans, Barcelona.
- MASCARÓ, J. (1983) *La fonologia catalana i el cicle fonològic*, Universitat Autònoma de Barcelona, Bellaterra.
- PACHÈS, P. (1999) *Improved Modelling for Robust Speech Recognition*, PhD thesis, Universitat Politècnica de Catalunya, Barcelona.
- PALMADA, B. (1994) *La fonologia del català. Els principis generals i la variació*, Universitat Autònoma de Barcelona, Bellaterra.
- RECASENS, D. (1991) *Fonètica descriptiva del català (Assaig de caracterització de la pronúncia del vocalisme i consonantisme del català al segle XX)*, Institut d'Estudis Catalans ("Biblioteca Filològica", XXI), Barcelona.
- RECASENS, D. (1993) *Fonètica i fonologia*, Biblioteca Universitària, Enciclopèdia Catalana.
- TELEVISIÓ DE CATALUNYA (1995) *El català a TV3. Llibre d'estil*, Edicions 62, Barcelona
- WELLS, J. (1987) "Computer-coded phonetic transcription", *Journal of the International Phonetic Association* 17, 2: 94-114.
- WELLS, J. (1989) "Computer-Code Phonemic Notation of Individual Languages of the European Community", *Journal of the International Phonetic Association* 19, 1: 31-54.
- WELLS, J. (1995) "Computer-coding the IPA: a Proposed extension of SAMPA". Available at: <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>.