SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

ESPRIT PROJECT 6819 (SAM-A0

Speech Technology Assessment
in Multilingual Applications

Source: Joaquim Llisterri
Universidad Autónoma de Barcelona

Tel 34 3 581 12 16
Fax 34 3 581 16 86

José B. Mariño
Universidad Politécnica de Cataluña

Tel 34 3 401 64 37
Fax 34 3 401 64 47

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

1. The Spanish adaptation of SAMPA

1.1. Phonetic and phonological inventories of Spanish

1.1.1. The phonetic inventory

The traditional description of the inventory of phonetic segments used by Peninsular Spanish speakers is found in Navarro (1918). His early descriptive work can be completed with the detailed list of allophones compiled by Canellada - Kuhlman (1987). The number of allophones quoted by these two sources amounts up to 20 vocalic elements and 43 consonantal segments. The causes of the allophonic variability according to these traditional sources can be summarized as follows:

Assimilations of place of articulation

• Interdental allophones of /t/, /n/ and /l/, the labiodental allophone of /m/, dental allophones of /n/, /s/ and /l/ and the palatal and velar allophones of /n/ are included in this category. The occurrence of these allophones is always conditioned by the place of articulation of the following segment.

• The vowels can be modified according to the following consonant: /a/ has a palatal allophone and a velar one; the quality of the other vowels is also changed by the following consonant in the same syllable, producing open and close varieties.

Changes in manner of articulation

• The approximant allophones of /b/, /d/ and /g/ and the affricate allophone of /y/ appear according to the character of the preceding consonant.

Devoicing and voicing

• /b/, /d/, /r/ and /g/ have devoiced allophones in syllable-final position before a voiceless consonant.

• /θ/ and /s/ have voiced allophones when they are followed by a voiced consonant.

Position in the syllable and syllabic type

• According to Navarro (1918) the Spanish vowels have close and open allophones depending on the structure of the syllable in which they appear; vowels tend to be closed in CV syllables and open in CVC syllables.

• The vowels /i/ and /u/ have allophonic variants -- known as semiconsonants or semivowels -- according to their nuclear or peripheral position in the syllable.

• The approximant allophones of /b/, /d/ and /g/ and the affricate allophone of /y/ are also conditioned by their position in the syllable

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

Position in the word

• Initial vs. non-initial position in the word is another factor that controls the appearance of the approximant allophones of /b/, /d/ and /g/ and the affricate allophone of /y/ .

• Traditional phoneticians such as Navarro (1918) distinguish lax allophones of the vowels depending on their position within the word.

Stress

• As well as the position in the word, the situation with respect to the main stress also results in lax allophones of the vowels. The affricate allophone of /y/ is also conditioned by the stress.

1.1.2. The phonological inventory

On the other hand, the inventory of phonological units proposed by classical authors such as Alarcos (1950) consists of 5 vowels and 19 consonants. The phonological segments identified by Alarcos are the following (transcribed according to IPA conventions):

Phoneme (IPA)	
p	voiceless labial plosive
b	voiced labial plosive
f	voiceless labiodental fricative
t	voiceless dental plosive
d	voiced dental plosive
θ	voiceless interdental fricative
tʃ	voiceless palatal affricate
y	voiced palatal fricative
s	voiceless alveolar fricative
k	voiceless velar plosive
g	voiced velar plosive
x	voiceless velar fricative
m	voiced labial nasal
n	voiced alveolar nasal
ɲ	voiced palatal nasal
l	voiced alveolar lateral
ʎ	voiced palatal lateral
ɾ	voiced alveolar tap
r	voiced alveolar trill

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

a	central open vowel
e	front mid vowel
i	front close vowel
o	back mid rounded vowel
u	back close rounded vowel

In order to be able to use a manageable number of units, but also to ensure a certain amount of phonetic detail, a compromise has been sought between the maximal number of allophones and the relatively short list of phonological units.

1.2. Statistical study of the occurrence of Spanish allophones

To arrive at such a compromise, a statistical study of the frequency of occurrence of the Spanish allophones has been undertaken. Since no data on the distribution of the allophones were available[1], it was decided to undertake a pilot experiment to evaluate the frequency of occurrence of the set of allophones described in the literature.

Three native Spanish speakers aged between 20 and 40 were interviewed by one experimenter for around one hour to obtain a large sample of speech. The interviewers restricted their interventions to the minimum, so that semi-spontaneous guided interviews were obtained. The recordings took place in an acoustically controlled environment using professional recording equipment An orthographic transcription was made, introducing punctuation according to prosodic, syntactic and semantic criteria. This transcription was the input of au automatic grapheme to allophone conversion programme, that generated a phonetic output with most of the allophones described in the literature. A sample of more that 100.000 segments was obtained, and the frequency of occurrence of each allophone, as well as other parameters, was computed.

1.3. Final inventory for Spanish

A final inventory was established by eliminating all the allophones with a frequency of occurrence below 0.10% in the corpus analyzed. Following this procedure, 31 segments were retained. The following table shows the IPA transcription for each allophone, its phonetic definition, the frequency of occurrence in the analyzed corpus and the frequency of occurrence quoted by Rojo (1991) when available.

IPA		% of occurrence in the corpus analyzed	% of occurrence according to Rojo (1991)
p	voiceless bilabial plosive	2.6	2.66

---

[1] Previously published studies were carried out considering only phonological segments (see. for example. Rojo (1991))

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

b	voiced bilabial plosive	0.45	2.66
t	voiceless dental plosive	4.63	4.48
d	voiced dental plosive	0.76	4.79
k	voiceless velar plosive	4.04	3.98
g	voiced velar plosive	0.11	0.95
m	voiced bilabial nasal	3.63	3.09
n	voiced alveolar nasal	7.02	6.99
ɲ	voiced palatal nasal	0.27	0.19
ŋ	voiced velar nasal	0.46	included in /n/
tʃ	voiceless palatal affricate	0.40	0.28
β	voiced bilabial approximant	2.47	included in /b/
f	voiceless labiodental fricative	0.51	0.68
θ	voiceless interdental fricative	1.53	1.68
ð	voiced dental approximant	3.20	included in /d/
s	voiceless alveolar fricative	6.95	7.58
z	voiced alveolar fricative	1.33	included in /s/
y	voiced palatal fricative	0.19	0.22
x	voiceless velar fricative	0.63	0.73
ɣ	voiced velar approximant	0.79	included in /g/
l	voiced alveolar lateral	4.25	5.08
ʎ	voiced palatal lateral	0.54	0.38
r	voiced alveolar trill	0.40	0.79
ɾ	voiced alveolar tap	4.25	5.67
i	front close vowel	4.29	7.5
j	voiced palatal approximant	2.60	included in /i/
e	front mid vowel	13.72	13.51
a	central open vowel	13.43	13.40
o	back mid rounded vowel	10.37	9.57
u	back close rounded vowel	1.98	3.16
w	voiced labial-velar approximant	1.35	included in /u/

Thus, the final inventory contains the 24 phonemes defined by Alarcos (1950) plus 7 segments traditionally considered allophones: the three approximant variants of the

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

voiced plosives [β ð ɣ], the voiced allophone of /s/ -- [z] -- the velar allophone of /n/ -- [ɲ]-- and the two semiconsonants or semivowels -- [j w] --.

1.4. Phonetic notation in Spanish using SAMPA

In our Spanish adaptation of SAMPA we have taken into account the proposals made by Wells (1989: 52-53), which are summarized below:

Approximant allophones of / b d g /

Since [D] (IPA [ð]) and [G] (IPA [ɣ]) already exist in SAMPA, only [B] is needed to represent the approximant [β]

Alveolar trill

The digraph [rr] can be used to represent the alveolar trill.

Affricate allophone of /y/

The affricate allophone of /y/ can be symbolized by [dZ] (IPA [dʒ]). However, this allophone has not been retained in our basic inventory due to its low frequency of occurrence.

Palatal fricative consonant

According to Wells (1989:52) the palatal fricative consonant /y/ can be considered an allophone of the semivowel [j]. Alarcos (1950 § 98) offers convincing arguments in favor of the phonological status of /y/ based on functional grounds and cites minimal pairs contrasting /y/ and the other consonants of the phonological system. His solution is widely accepted in the literature on Spanish phonetics and phonology and, moreover, it does not seem to be counterintuitive with regard to native speakers' intuitions. It is widely accepted that the phoneme /y/ can be realized as a fricative, as an approximant, and also as an affricate under certain conditions. Following Wells' (personal communication) suggestion, this phoneme will be represented in SAMPA by the digraph /jj/.

The following table summarizes the set of symbols that can be used in SAMPA to transcribe the phonemes and allophones of Spanish selected according to the previously described criteria.

IPA	SAMPA		Example	Transcription
p	**p**	voiceless bilabial plosive	**p**ala	"pala
b	**b**	voiced bilabial plosive	**b**ala	"bala
t	**t**	voiceless dental plosive	**t**ala	"tala
d	**d**	voiced dental plosive	**d**ar	dar
k	**k**	voiceless velar plosive	**c**ala	"kala
g	**g**	voiced velar plosive	**g**ala	"gala

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

m	**m**	voiced bilabial nasal	**m**ala	"mala
n	**n**	voiced alveolar nasal	**n**ada	"naDa
ɲ	**J**	voiced palatal nasal	ca**ñ**a	"kaJa
ŋ	**N**	voiced velar nasal	ho**n**go	"oNgo
tʃ	**tS**	voiceless palatal affricate	**ch**ico	"tSiko
β	**B**	voiced bilabial approximant	la**v**a	"laBa
f	**f**	voiceless labiodental fricative	**f**also	"falso
θ	**T**	voiceless interdental fricative	**z**ona	"Tona
ð	**D**	voiced dental approximant	ca**d**a	"kaDa
s	**s**	voiceless alveolar fricative	**s**ala	"sala
z	**z**	voiced alveolar fricative	de**s**de	"dezDe
y	**jj**	voiced palatal fricative	a**y**er	a"jjer
x	**x**	voiceless velar fricative	**j**amón	xa"mon
ɣ	**G**	voiced velar approximant	la**g**o	"laGo
l	**l**	voiced alveolar lateral	**l**a	la
ʎ	**L**	voiced palatal lateral	**ll**ana	"Lana
r	**rr**	voiced alveolar trill	ca**rr**o	"karro
ɾ	**r**	voiced alveolar tap	ca**r**o	"karo
i	**i**	front close vowel	t**i**la	"tila
j	**j**	voiced palatal approximant	lab**i**o	"laBjo
e	**e**	front mid vowel	t**e**la	"tela
a	**a**	central open vowel	t**a**l	tal
o	**o**	back mid rounded vowel	t**o**do	"toDo
u	**u**	back close rounded vowel	t**u**l	tul
w	**w**	voiced labial-velar approximant	ag**u**a	"aGwa

If there is a need to represent other allophones not present in the set of segments described, the following SAMPA symbols are available:

IPA	SAMPA		Example	Transcription
dʒ	**dZ**	voiced palatal affricate	con**y**ugal	kondZu"Gal

2. Automatic phonetic transcription for Spanish: generating SAMPA representations from orthographic representations

2.1. Grapheme to allophone correspondences

In order to produce an automatic transcription the correspondences between the graphemes and the SAMPA symbols have to be established. The following table summarizes some of the main correspondences that has been taken into account to design the transcription algorithm.

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

grapheme	rules for the transcription to SAMPA	examples
<a>	**a**	
<b>	after a pause: **b** after <m> or <n>: **b** other cases: **B**	com**b**a" "kom**b**a la**b**io: "la**B**jo
<c>	followed by <e> or <i>: **T** in word final position followed by <l> or <r>: **G** followed by <b>, <d>, <g> (preceding <a>, <o>, <u>), <m>, <n>, <ñ> or <v>: **G** other cases **K**	**c**elo: "**T**elo a**c**né: a**G**"ne ta**c**to: "ta**k**to **c**oro" "**k**o**r**o te**c**la: "te**k**la
<ch>	**tS**	**ch**elo: "**tS**elo
<d>	after a pause: **d** after <l>, <m> or <n>: **d** other cases: **D**	cal**d**o: "kal**d**o co**d**o: "ko**D**o
<e>	**e**	
<f>	**f**	co**f**ia: "ko**f**ja
<g>	after a pause and followed by <r>, <l>, <a>, <o> or <u>: **g** after <m> or <n> and followed by <a>, <o> or <u>: **g** followed by <i> or <e>: **x** other cases: **G**	ton**g**o: "ton**g**o **g**enio: "**x**enjo ti**g**re: "ti**G**re la**g**o: "la**G**o
<h>	in word-initial position followed by <ie>: **jj** other cases: no sound	**h**ierba: "**jj**er**B**a **h**alo: "alo
<i>	in nuclear position in the syllable: **i** in non nuclear position in the syllable: **j**	t**i**po: "t**i**po c**i**elo: "T**j**elo
<j>	**x**	**j**arana: "**x**arana
<k>	**k**	**k**iosko: "**k**josko
<l>	**l**	**l**ote: "**l**ote
<ll>	**L**	ta**ll**o: "ta**L**o
<m>	**m**	ar**m**a: "ar**m**a
<n>	followed by <p>, <b>, <v>, <m> or <f>: **m** other cases: **n**	á**n**fora: "a**m**fora co**n**o: "ko**n**o
<ñ>	**J**	u**ñ**a: "u**J**a
<o>	**o**	
<p>	**p**	**p**erro: "**p**erro
<q>	always followed by <u>: **k**	**q**ueso: "**k**eso
<r>	in word-initial position: **rr** preceded by <l>, <n> or <s>: **rr** other cases: **r**	**r**ama: "**rr**ama hon**r**a: "on**rr**a a**r**pa: "a**r**pa t**r**ampa: "t**r**ampa pe**r**a: "pe**r**a amo**r**: a"mo**r**

SAM-A
Spanish adaptation of SAMPA and
automatic phonetic transcription

<rr>	**rr**	carro: "ka**rr**o
<s>	**s**	ra**s**go: "**rr**a**s**Go ca**s**a: "ka**s**a tra**s**to: tra**s**to
<t>	in syllable-final position: **D** other cases: **t**	a**t**leta: a**D**"leta **t**oro: "**t**oro
<u>	without diaresis preceded by <g> or <q>: no sound in non nuclear position in the syllable: **w** in nuclear position in the syllable: **u**	q**u**eso: "keso cig**ü**eña: Ti"G**w**eJa l**u**jo: "l**u**xo
<v>	after a pause: **b** after <m> or <n>: **b** other cases: **B**	con **v**elo: kom "**b**elo cal**v**o: "kal**B**O
<w>	in foreing words: **Gu**, **gü** or like a <v>	**w**hisky: "**gw**iski ki**w**i: "ki**B**i
<x>	in non word-initial position and followed by a vowel: **Gs** other cases: **s**	e**x**amen: e**G**"**s**amen e**x**terno: e**s**"terno
<y>	in initial position of a syllable with two or more sounds: **jj** other cases: it will be processed like a <i>	**y**unque: "**jj**unke cón**y**uge: "kon**jj**uGe dos **y** dos: dos **i** "Dos mu**y**: mw**i**
<z>	**T**	**z**ar**z**a: "**T**ar**T**a ti**z**ne: "ti**T**ne

3. References

ALARCOS, E. (1950) *Fonología española.* Madrid: Gredos (Biblioteca Románica Hispánica, Manuales 1), 1965 4a ed. aumentada y revisada.

CANELLADA, M. J. - KUHLMAN MADSEN, J. (1987) *Pronunciación del español. Lengua hablada y literaria*. Madrid: Castalia.

NAVARRO TOMÁS, T. (1918) *Manual de pronunciación española.* Consejo Superior de Investigaciones Científicas: Madrid, Instituto Miguel de Cervantes (Publicaciones de la Revista de Filología Española, III). 21ª edición, 1982.

ROJO, G. (1991) " Frecuencia de fonemas en español actual ", in BREA, M.- FERNANDEZ REI, F. (Coord) *Homenaxe ó profesor Constantino García*. Santiago de Compostela: Universidade de Santiago de Compostela. Servicio de Publicación e Intercambio Científico. Pp. 451-467.

WELLS, J. C. (1989) " Computer-coded phonemic notation of individual languages of the European Community ", *Journal of the International Phonetic Association* 19,1: 31-54