

DESIGN AND EVALUATION OF A PHONOLOGICAL PHRASE PARSER FOR SPANISH TEXT-TO-SPEECH

Helen E. Karn
Georgetown University

ABSTRACT

This paper presents and evaluates a phonological phrase parser for a Spanish text-to-speech system. The parser consists of three stages: 1) lexical lookup, using a small dictionary (428 words); 2) preliminary phrase boundary placement, using a modification of Liberman and Church's (1992) function group parser; and 3) readjustment of phrase boundaries, using syllable count and punctuation. A corpus of 382 hand-parsed sentences (1,691 phrases) was used to evaluate the parser. The parser generated almost the same number of phrases (1,692) as the hand-parsed sentences with 70% (1,186) agreement. Suggestions for improving the parser's performance include the expansion of the verb lexicon, performing simple morphological analysis for verbs, and relaxing the syllable count in phrases before verb forms.

1. INTRODUCTION

An important step in assigning intonation contours in a text-to-speech (TTS) system is determining the units in which the intonation contours will apply. These units are referred to here as phonological phrases. One method of determining phonological phrases is the presence of punctuation: assign an intonation contour to each group of words which falls between punctuation marks. However, this method is less than satisfactory for sentences which contain little or no punctuation. Another method is to assign an intonation contour to surface syntactic constituents, but phonological phrases do not always coincide with syntactic phrases^{1,2}. Furthermore, for unrestricted texts, it is more important to have a broad syntactic analysis than a deep analysis of a limited domain³.

2. PARSER DESIGN

The parser operates at the sentence level on texts which contain conventional Spanish orthography and punctuation. The sentence is defined here as a string of text bound by major punctuation (periods, quotation marks, question marks, and exclamation marks), which is indicated in the examples here by two forward slashes (//).

2.1 Lexical Lookup

Each word of the input sentence is first assigned to one of four categories: Function Word, Auxiliary Word, Verb, or Content Word. The category of Content Word is assigned to input words which do not match any of the entries in the Function, Auxiliary, or Verb lexicons. (See Karn (1996) for the lexical entries.)

Function Words. The function word lexicon (136 entries) consists of the closed classes of determiners, prepositions, conjunctions, disjunctions; and personal, possessive, and interrogative pronouns. The inclusion of conjunctions and disjunctions in the Function Word lexicon differs from a previous approach⁵, which uses these items to generate major and minor phrase boundaries. Only major phrase boundary placement is considered here.

Auxiliary Words. The auxiliary word lexicon (25 entries) consists of forms of the auxiliary verb *haber* 'have' and words which begin verb phrases, such as negators (*no, nunca*), reflexive pronouns (*me, te, se, os, nos*), and indirect pronouns (*le, les*).

Verbs. The verb lexicon (267 entries) consists of the finite verb forms of the 2,000 most frequent words in a corpus of modern written Mexican Spanish⁴. Nonfinite verb forms (infinitives, participles) do not appear in the verb lexicon and are considered Content Words, because they do not function as predicates by themselves. For example, infinitives occur in the subject function, such as *cantar* 'to sing' in the sentence *El cantar es divertido* 'Singing is fun'. Present and past participles occur after forms of the auxiliary verb *haber*, so it is redundant to list them in the verb lexicon.

Lexical lookup for the sentence *No es una carta de Chicago* 'It isn't a letter from Chicago' is shown in (1).

1.	No	es	una	carta
	not	is	a	letter
	AUX	VERB	FUNCTION	CONTENT
	de	Chicago		
	from	Chicago //		
	FUNCTION	CONTENT		

2.2 Preliminary Phrase Boundary Placement

Next, a preliminary phrase boundary is placed between pairs of words according to the scheme shown in Figure 1. The placement of the preliminary phrase boundaries (Figure 1) is a modification for Spanish of Liberman and Church's (1992) function group (f-group) parser for English.

For the sentence in (1), boundaries are placed after *es* ‘is’ and *carta* ‘letter’, which produces the three preliminary phrases shown in (2).

2. (No es)
 AUX VERB
- (una carta)
 FUNCTION CONTENT
- (de Chicago) //
 FUNCTION CONTENT

As Figure 1 shows, neither Function Words nor Auxiliary Words trigger a break between them and words from any of the four lexical categories. However, these categories are kept separate in order to identify verb phrases. (See Section 3.3 for further discussion on labeling verbs and verb phrases.)

Category of Left Word	Category of Current Word			
	Function Word	Content Word	Auxiliary Word	Verb
Function Word	No Break <i>de los</i> ‘of the’	No Break <i>de María</i> ‘of María’	No Break <i>y ha</i> ‘and has’	No Break <i>lo ví</i> ‘it I-saw’
Content Word	Break <i>María / pero</i> ‘María but’	No Break <i>casa bonita</i> ‘house pretty’	Break <i>María / ha</i> ‘María has’	Break <i>María / sabe</i> ‘María knows’
Auxiliary Word	No Break <i>me lo</i> ‘to-me it’	No Break <i>ha puesto</i> ‘has put’	No Break <i>no lo</i> ‘not it’	No Break <i>no es</i> ‘not is’
Verb	Break <i>sabe / el</i> ‘knows / the’	Break <i>tiene / tres</i> ‘has / three’	Break <i>sabe / no</i> ‘knows / not’	Break <i>ví / tiene</i> I-saw / has

Figure 1. Preliminary phrase boundary placement.

2.3 Readjustment of Boundaries

Last, the parser readjusts the preliminary phrases by combining two or more phrases. A phrase which 1) does not end in punctuation and 2) contains 10 or fewer syllables is combined with the following phrase until a punctuation mark is reached or until the number of syllables is greater than 10. (Both stressed and stressless syllables are included in the syllable count.) This approach differs from a previous approach⁵, which counts syllables in the entire sentence, rather than in the preliminary phrases. To obtain a syllable count, the graphemic text is first converted to phonemes and then syllabified. (See Karn 1996 for details.) For the purposes of this paper, the graphemic versions of the phonological phrases are given.

The three preliminary phrases in (2) contain a total of 10 syllables: *no es* (two syllables), *una carta* (four), *de Chicago* (four). Because there is no punctuation until the end of the sentence, the three preliminary phrases of (2) are combined into one phonological phrase, shown in (3).

3. (No es una carta de Chicago) //

3. PARSER EVALUATION

3.1 The Corpus

A set of 382 sentences (1,691 phrases) from 34 written texts^{6,7,8,9}, whose phonological phrase boundaries were marked by expert phonologists, was used to evaluate the parser. The texts were wordprocessed according to standard Spanish orthography and stored as ASCII text files.

3.2 Classification of Boundaries

The boundaries produced by the parser were compared to those marked by the expert and classified as *Correct*, if they matched the expert’s boundaries, or *Incorrect*, if they did not. The category of *Incorrect* was subdivided into *Marked* and *Not Marked*. A *Marked* boundary was one placed by the parser, but not the expert. A boundary that was *Not Marked* was placed by the expert, but not the parser. The sentence shown in (4) illustrates these three groups.

4. Phrases produced by parser for Sentence 9 of *La Gloria de Don Ramiro*⁶.

Todos presentían que aquella cabeza
 (Correct)

no estaba segura sobre el soberbio cogote
 (Correct)

y esperaban por momentos alguna catástrofe;
 (Correct)

pero el hidalgo / [demostraba importársele
(Not Marked)

una higa] / de la delación y del riesgo,
(Not Marked)

perorando aun con más vivo coraje
(Correct)

[cuando se hallaban presentes el señor corregidor don
Alonso
(Marked)

de Cárcamo] / o el fraile dominico
(Not Marked)

[en quien todos sospechaban un espía
(Marked)

del Santo Oficio y del monarca.]
(Correct)

3.3 Results.

The parser generated almost the same number of boundaries as the experts: 1,692 vs. 1,691, respectively. The placement of the boundaries by the parser agreed with those of the experts in 1,186 (70%) of the phrases.

Of the incorrect boundaries, the majority (356; 70.5%) were *Not Marked* by the parser. In these cases, the parser placed preliminary boundaries which were not placed by the experts or it readjusted (combined) phonological phrases boundaries where the experts had not.

Placement of preliminary boundaries: verbs and content words. Recognizing verb forms was difficult for the parser: 27.5% (98) of the words occurring before *Not Marked* boundaries were verbs. Because the verb lexicon contains only a small number of finite verb forms, verbs were mislabeled as Content Words. Because the parser does not separate groups of Content Words (see Figure 1), preliminary boundaries were not marked in these cases.

Example (5) shows the result of this mislabeling. The verb *surgen* ‘soar’ is not in the verb lexicon, so it was labeled a Content Word and grouped with the surrounding Content Words, *cables* ‘cables’ and *rígidas* ‘rigid’.

5. Las torrecillas metálicas de los *cables surgen rígidamente*
the towers-small metal of the *cables / soar rigidly*
‘The small metal towers of the cables / soar rigidly’

Incorrect combining of preliminary phonological phrases. In readjusting the preliminary boundaries, the parser considers only the number of syllables in the phrase and the presence of

punctuation. Therefore, the parser combines some preliminary phonological phrases which should remain as separate phrases.

In example (6), *era* ‘it was’ was correctly labeled Verb, so a preliminary boundary was placed between *era* and the previous Content Word *pregunta* ‘question’. Boundary readjustment then combined these phrases because the number of syllables in the first two phrases (*dirigírselo* and *una pregunta*) was nine.

6. (Dirigírselo) 4 syllables
direct-to-him

(una pregunta) 5 syllables
a question

(era como abrir)
was like open-inf.

(Dirigírselo una pregunta era como abrir) /
‘Asking him a question was like opening’

In the case of boundaries which were *Marked*, the parser had difficulty because of its lack of semantic information, which affected the following areas.

Idiomatic expressions. Idiomatic expressions function as single semantic units but may contain more than one syntactic phrase. In (7), the idiomatic expression *importársele una higa* ‘not care a bit’ was incorrectly divided by the parser into two preliminary phrases: the verb *importársele* ‘to care’ and the noun phrase *una higa* ‘a bit (lit. a fig)’. The syllable count of the first phrase was greater than ten, so it was not combined with the second phrase. (In addition, the verb *demostraba* ‘demonstrated’ was labeled as Content Word.)

7. pero el hidalgo / [demostraba *importársele*
(Not Marked)

una higa] / de la delación y del riesgo,
(Not Marked)

The preposition *de* ‘of/from’. This preposition, which is the most frequent function word in Spanish⁴, posed another problem for the parser. The proper name *Don Alonso de Cárcamo* in (4) was incorrectly divided into the two groups *Don Alonso* and *de Cárcamo*, because *Alonso* was labeled a Content Word and *de* a Function Word. As in (7), the two groups were not combined because the syllable count of the first phrase was greater than ten syllables.

Adjectival phrases with *de* were also incorrectly divided by the parser, as in the noun phrase *a wooden bench* of (8):

8. Martín Marco se sienta en *un banco / de madera* y ...
Martin Marco sits down on *a bench / of wood* and ...
‘Martín Marco sits down on a wooden bench and ...’

4. SUGGESTIONS FOR IMPROVEMENT

The parser can be modified in several ways to reduce the number of ill-formed preliminary boundaries.

First, the verb lexicon could be expanded to include more finite verb forms: verbs such as *comenzaban* 'they were beginning', *demostraba* 'It showed', and *empezaron* 'they started' are not currently in the lexicon. In addition, a simple morphological component could be added to check words labeled Content Words for verb suffixes. These suffixes should be unambiguous forms such as *-íamos* and *-ieron*; the suffix *-ía* could erroneously label proper names such as *María* and *Lucía* as verbs. Furthermore, words following Auxiliary Words could be marked as Verb, rather than Content Word to reduce long strings of Content Words.

Second, lexicons of idiomatic expressions and proper names containing *de* could be added. Idiomatic expressions containing verbs (e.g. Example 7) should undergo simple morphological parsing as well.

If the verbs of the input sentences can be reliably detected (e.g. by using a large lexicon of finite verb forms), then a Case Grammar/Thematic Relations approach could be incorporated in the parser. This approach would focus on the verb as the central element in the sentence and look for the roles associated with the verb, such as Benefactor, Object, and Location. This approach, however, would require much more extensive parsing (and lexicons) than the scheme presented here. In addition, it is not immediately known whether such an approach would greatly improve the parser's performance.

Third, to reduce the number of boundaries which are inappropriately combined, the syllable limit of phrases immediately preceding verbs could be reduced to less than ten syllables. However, what this new limit would be and whether this would result in an overgeneration of phonological phrases would have to be tested.

5. REFERENCES

1. Bachenko, J., and Fitzpatrick, E. "A computational grammar of discourse-neutral prosodic phrasing in English." *Computational Linguistics*, 16, 3:155-70, 1990.
2. Karn, H. E. *Aspects of Phonetics and Phonology in a Text-to-Speech System for Spanish*. Ph.D. dissertation, Georgetown University, 1996.
3. Liberman, M. Y., and Church, K. W. "Text analysis and word pronunciation in text-to-speech synthesis." *Advances in speech signal processing*, ed. by S. Furui and M. M. Sondhi, 791-831. New York, Marcel Dekker, 1992.
4. Chandler-Burns, R., et al. *SPANISH 1992: Corpus-based analysis of present-day Spanish for medical purposes*. FTP: wintermute.fullerton.edu (pub/research/chandler); rchandler@ccr.dsi.uanl.mx. Fax: (528) 348-5477; Prof. Robert M. Chandler-Burns, Autonomous University of Nuevo León, College of Medicine, Apdo. Postal 1563, CP 64000, Monterrey, Mexico, 1992.
5. Karn, H. E. "An algorithm for generating phrase boundaries for the automatic assignment of prosodic contours in a text-to-speech system for Spanish." Paper 4aSP7, 126th Meeting of the Acoustical Society of America, Denver, CO, October 1993.
6. Canellada, M. J. *Antología de textos fonéticos*. 2nd edn. Madrid, Gredos, 1972.
7. Canellada, M. J., and Madsen, J. K. *Pronunciación del español: Lengua hablada y literaria*. Madrid, Castalia, 1987.
8. Navarro Tomás, T. *Manual de pronunciación española*. 14th edn. Madrid, Consejo Superior de Investigaciones Científicas, 1968.
9. Quilis, A., and Fernández, J.A. *Curso de fonética y fonología españoles*. 10th edn. Madrid, Consejo Superior de Investigaciones Científicas, 1982.