



AUTOMATIC ASSIGNMENT OF LEXICAL STRESS IN ITALIAN

Philippe Martin

University of Toronto
Centre National d'Etude des Télécommunications (France)

ABSTRACT

An automatic method to assign lexical stress in Italian from a written text is presented here. This method is based on specific morphological properties of the Italian morphemes, which can only be stressed on the last or the penultimate syllable. Using a morphological analyzer and a morpheme database, the assignment program attempts to analyze each entry into a morpheme followed by one or more suffixes and flexions, which can contain a stressable syllable. The resulting stressed syllable is determined by applying a simple final stress rule to these elements.

1. INTRODUCTION

It is often difficult for a student of a free accent foreign language to master word stress, for which no stress rules are known. The same applies of course to a computer program which would have to position the lexical stress automatically.

In the case of speech synthesis of Italian from text, it is necessary to determine the position of the stressed syllable in the word in order to generate appropriate vocalic durations and pitch contours [5].

An automatic method to position word stress is presented here, based on morphological analysis of the word and the use of an Italian morpheme database.

2. THE PROBLEM

Italian is a free stress language, i.e. the stress position is not determined by a simple rule as in French (on the last pronounced syllable) or in Finnish (on the first syllable).

In Italian nouns, adjectives, pronouns and adverbs, stress can fall on the last

syllable (virt'ù, caff'è), on the penultimate (acc'ento, cont'ino), the antepenultimate (tel'efono, c'elebre), and even on the 4th syllable from the end of the word (c'austico).

Due to the adjunction of suffixes, verbs can be stressed from the last to the 6th syllable from the end. For example,

oxyton case: amer'ò (1st pers. sing. future)
"I will love"

paroxyton case: am'are (infinitive)
"to love"

proparoxyton case: pr'endilo (imperative,
"take it")

4th position from the end: f'abbricalo
(imperative, "fabricate it")

5th position from the end: f'abbricamelo
(imperative, "fabricate it for me")

6th position from the end: f'abbricamicelo
(imperative, "fabricate it for me to him")

More complicated situations occur with homographs, belonging to distinct grammatical categories, or worse (for a computer program), to the same grammatical categories.

In the often quoted example

Sono cose che capitano, capitano

("those are things that happen, captain")

the first capitano is a verb and is stressed on the 4th syllable from the end (c'apitano), whereas the second is a noun and is stressed on the penultimate syllable (capit'ano).

On the other hand, the homographs pr'incipi ("princes"), princ'ipi ("principles"), or t'urbine ("whirlwind") and turb'ine ("turbines") belong to the same grammatical categories. In those cases, ambiguity may be resolved with the syntactic context, as with the presence of an article (il t'urbine vs. le turb'ine, plural of la turbina).

The position of stress can thus be related to lexicon. The Italian terminology for the 6 possible cases is:

position n (last): tronco
 position n-1 : piano
 position n-2 : sdrucchiolo
 position n-3 : bisdrucchiolo
 position n-4 : trisdrucchiolo
 position n-5 : quadrisdrucchiolo

but only the first case (tronco) is marked graphically nowadays (virtù, caffè).

The position of stress varies sometimes according to the dialect, or with poetic effects (t'enebra vs. ten'ebra, poetic form) [8].

3. A STATISTICAL APPROACH

Speech synthesis applications are frequently elaborated by signal analysis specialists rather than linguists. It can thus be expected that automatic stress assignment implemented in this case will not use any morphological insight.

This is effectively the case in the speech synthesis system for Italian of the CSELT [2],[3].

Statistical observation made on about 8000 words shows that 78% are stressed on the penultimate syllable [2]. With only one rule always assigning stress on the penultimate, the error rate is already of 22%

The approach taken by the CSELT is based on the correlations observed between orthographic trigrams (sequences of 3 graphemes) and the location of the stress within the word. These correlations are implemented in about 250 rules used by an ATN automata operating from the end of the word.

For instance, the rules pertaining the the grapheme i are

	4	3	2	1
Stressed Element				
	+I,O,P,R,U 2	S	I	A,E

Sequences such as -ISIA, -ISIE, -OSIA, -OSIE, -PSIA, -PSIE, etc, will then be stressed on the penultimate syllable.

The appropriate ranking of these rules and the extensive use of lists of exceptions allows the system to reach a correct positioning rate of about 97%, outside explicit use of any phonetic or morphological properties, although careful examination of these rules would reveal the rediscovery of such properties. The correct positioning rate deteriorates somewhat for verbal forms and when homographs are present in the text.

4. A PHONOLOGICAL-PHONETIC APPROACH

From a linguistic point of view, it may seem reasonable to think that stress is linked somewhat to the phonological structure of the syllable. It would then be possible to establish contextual rules to determine the stressed character of a syllable from its phonetic and/or phonological structure.

This approach was used by Delmonte [1], and implemented in a speech synthesis system. The problems related to this method are due to the large number of rules (and the large number of exceptions) needed to obtain satisfactory performances.

After suitable orthographic-phonetic conversion, the rules act on the nature of groups of 3 elements inside the syllable and based on the vowel inside that syllable. For instance, the /i/ rule could be something like:

If /i/ belongs to the penultimate syllable, the word is

- proparoxytonic if /i/ is followed by /t d l m k t / and if the word does not belong to the list of exceptions;

- proparoxytonic if /i/ is followed by /l m n y t / and the word is a verb followed by a clitic;

- proparoxytonic if /i/ is followed by /g r n t/ and the word is a verb of the first group in -inare, -igare, -itare, or a noun or an adjective belonging to an exception list;

- paroxytonic in all the other cases.

Then we have the rules of /i/ with a left context (gi, mi, di, etc.), which have their own lists of exceptions, etc.

5. A MORPHO-PHONETICAL APPROACH

In [4], phonetic rules work in conjunction with morphological rules pertaining to the stressability of suffixes (considered as morphological units) For example, suffixes -illa, -esse are always stressable (they can be stressed) (dist'illo, profet'essa), whereas suffixes ido, -bile are non stressable (t'imido, sens'ibile).

In these cases too it is necessary to add lists of exceptions to obtain correct localisation of stress.

6. A MORPHOLOGICAL APPROACH

We can easily foresee that the development and the maintenance of stress rules (essentially of a phonetic nature in [2]) is difficult [7]. Their major drawback resides in their interdependence, the revision of one of them often provoking modifications in others and in their list of exceptions.

Since the advent of inexpensive computer equipment, a brute force approach may be considered, which would consist of entering a very large number of forms together with the position of the stressed syllable and the grammatical category. This latter characteristic would allow the treatment of homomorphic cases (which must be resolved in the orthographic-phonetic transcription as well).

This kind of approach, although used for other languages (and in particular for French), seems somewhat too difficult to handle. Indeed, to the 50,000 basic forms, it would be necessary to add 3 supplementary entries for the adjectives, about 50 forms for verbs, etc.

Aside from the use of a relatively small number of (difficult to modify) rules and the elaboration of a 250,000 plus lexicon, we propose another approach based on a relatively unknown property of lexical stress in Italian. This approach proceeds by morphological analysis of each entry into its morphological root (lexeme) and its suffixes and flexions (morphemes).

The method is based on the mechanism of stress assignment in Italian rather than on properties of syllables or suffixes. This mechanism (cf. [12], [13]), analyzes the word in its constituent morphological elements, a lexeme, followed (preceded) by one or more morphemes.

Morphemes, and in particular suffixes, can either be stressable or unstressable. For example, -illa is stressable (can receive the lexical stress of the word), and -ido is unstressable (cannot contain the word stress).

Depending on their nature, i.e. the lexeme they determine, homographic suffixes can be stressable or unstressable. For instance, -ino, masculine singular diminutive suffix, is stressable (piccol'ino), but -ino, 3rd person subjunctive plural suffix is unstressable ('amino). By the same token, -o unstressable is the 1st person indicative present verbal suffix, and -o stressable is the 3rd person past absolute singular verbal suffix.

Lexemes are always stressable (but not necessarily stressed), either on their last syllable (ar'en-, in ar'ena), or on their penultimate syllable (f'abbric-, in f'abbrica). According to [13], 82% of lexemes are stressable on the last syllable and only 18% on the penultimate.

The stress rule simply specify that, in a sequence

(prefix)+(lexeme)+(suffix)+...+(suffix)

in which one or more elements are stressable, the last stressable unit (lexeme or suffix) will determine the position of the word stress.

For example, the derivatives of 'oper, stressable on the last syllable, receive their word stress as follows:

'opera : 'oper + -a (unstressable suffix)
"opera"

oper'oso : 'oper + -'os- (stressable suffix) + -o "hard worker"

oper'etta : 'oper + -'etta (stressable suffix) "operetta"

operosit'a: 'oper + -'os- + -it'a (stressable suffix) "industriousness"

By the same token, we have:

turb'ina : turb'in + -a "turbine"

t'urbine : t'urbin + -e "whirlwind"

The position of stress in the lexeme allows to distinguish between these two derivatives of the same Latin word.

In some cases, the word can be analyzed into two lexemes.

The two lexemes can be either interdependent (they cannot appear alone) or one or both of them are independent. In the first case the **first** lexeme will receive the word stress [18]:

tel'efono : tel'e + f'ono "telephone"

s'incrono : 'sin + cr'ono "synchronous"

If the last lexeme is independent or if both lexemes are independent, the last stressable unit (lexeme or suffix) will determine the position of the word stress.

aeron'ave : a'ero + n'av + -e "airplane"

alisc'afo : 'ali + sc'af + -o "hydrofoil"

Automatic assignment of word stress, which, in complete forms, can be located between the last and the 6th syllable from the end, can thus be handled by a morphological analysis and by identification of the stressable property of the units discovered by this analysis through the use of an appropriate lexicon.

Since many suffixes are homomorphic, this process requires proper matching of suffix grammatical categories to the corresponding categories of the lexemes.

Schematically, the complete process is based on the search into two databases:

- a lexicon of lexemes containing their grammatical category, the position of their stressable syllables well as their independent or dependent nature;

- a lexicon of morphemes indicating the position of the stressable syllable.

The use of the first lexicon allows the elaboration of possible analysis hypothesis. In the general case, more than one morphological root can be associated with the analyzed entry:

capitano:

capit'an- (noun) + -o

c'apit- (verb) + -ano

The identification of last component of the first process is done by consultation of a second lexicon of morphemes:

-o : morpheme masc sing, unstressable

-a-: verbal flexion (plural), unstressable

-no: verbal flexion (3rd pers.), unstressable

Applying the final stressable syllable rule determines the resulting position of the word stress.

The two possible analyses of the example are then:

capit'an-o : noun ("captain")

c'apit-an-o : verbal form ("happen")

The context allows disambiguation at the sentence level if necessary.

In

Sono cose che c'apitano, capit'ano

("these are things that happen, captain")

the punctuation sign in front of the second capitano indicates its category as a noun.

7. IMPLEMENTATION

Automatic assignment of lexical stress in Italian by morphological analysis requires:

- a lexicon of about 50,000 entries;
- a lexicon of about 300 lexemes (suffixes and flexions)
- an analysis program allowing a fast consultation of these two databases.

The lexeme acquisition program allows easy entry of the lexeme form and the marking of its properties (grammatical category, gender, number, person, morphological type, mode, tense, etc.). In order to mark the stressable syllable, the system automatically presents to the operator a form with non stressable suffixes and/or flexions, by consultation of the appropriate lexicon. The resulting word is then necessarily stressed on the root and marking of the stressable syllable can be done.

The final implementation of the system includes disambiguation of homomorphic cases based on the occurrences of grammatical categories, as well as stress clash rules [6],[11].

REFERENCES

- [1] R. Delmonte (1981) "L'accento di parola nella prosodia dell'enunciato dell'italiano standard", Studi di Grammatica Italiana, Vol. X, 351-394.
- [2] S. Sandri and E. Vivalda (1981) "Automatic Stress Assignment for Italian Text-to-Speech Synthesis", CSELT Rapporti Tecnici, Vol. VIII, No 3, juin 1981, 213-216.
- [3] S. Quazza and E. Vivalda (1987) "Contextual Syntactic Analysis for Text-to-Speech Conversion", Proc. ICASSP 87, 389-392.
- [4] O. Profili (1987) "L'accent et sa prévisibilité", Rapport Syntalit/ Italien, CNET-Lannion.
- [5] Ph. Martin (1978) "L'intonation de la phrase en italien", Studi di Grammatica Italiana, VIII, Acc. della Crusca, Firenze.
- [6] M. Nespore and I. Vogel (1979) "Clash avoidance in Italian", Linguistic Inquiry, X, 3, 467-482.
- [7] E. Laporte (1986) "Application de la morpho-phonologie à la production automatique de textes phonétiques", Actes GALF Symposium Lexique, 215-227.
- [11] Ph. Martin and O. Profili (1987) "Accent de mot et structure syntaxique en italien", Information-Communication, U. of Toronto, 15-26.
- [12] P. Garde (1968) "L'accent", PUF, Paris.
- [13] P. Antonetti and M. Rossi (1970) "Précis de Phonétique Italienne. Synchronie et Diachronie", Aix-en-Provence, 356p.
- [18] M. Rossi (1979) "Le cadre accentuel et le mot en italien et en français", in Problèmes de prosodie, Studia Phonetica 17, Léon and Rossi, éd. Didier, Montréal, 9-22.

This research is part of a CNET contract No 88 1B 108.