A SYSTEM OF ELECTRONIC DICTIONARIES OF PORTUGUESE

SAMUEL ELEUTÉRIO

Instituto Superior Técnico
Av. Rovisco Pais, P-1096 Lisboa Codex, Portugal

ELISABETE MARQUES RANCHHOD

Universidade de Lisboa
Cidade Universitária, P-1699 Lisboa Codex, Portugal

HELENA FREIRE

Instituto Politécnico de Portalegre
Praça do Município, P-1300 Portalegre, Portugal

JORGE BAPTISTA

Universidade do Algarve
Gambelas, P-8000 Faro, Portugal

SUMMARY

In this paper we present the state of development of DIGRAMA, a system of electronic dictionaries and grammars of Portuguese. We describe briefly the general characteristics of the system, concentrating then on the analysis of two dictionary modules: DIGRAS and DIGRAF. The former is a dictionary of simple words, the latter is a dictionary of inflected forms. To each lexical entry of DIGRAS has been associated a code corresponding to the formal description of the morphological properties of the entries. This information constitutes the basis for a program of automatic generation of inflected forms. DIGRAF is made up of all the inflected forms from the words of DIGRAS.

January 1993

0. Presentation

DIGRAMA is a system of electronic dictionaries and grammars of Portuguese, *ie* lexical and grammatical data bases specifically build to be used by computers in natural language processing.

The dictionary modules contain the words and the formal description of their morphological properties. The grammars correspond to the matrices of the lexicon-grammar of Portuguese, where free and frozen constructions of verbs, adjectives, predicative nouns and adverbs are described and formalized (see References).

In this paper we refer to two dictionary modules: a dictionary of simple words and a dictionary of inflected forms. These dictionaries allow: (i) to recognize a given word form both canonical and inflected; (ii) to include a given form into the grammatical class(es) it belongs to; (iii) to show the inflected forms associated to a word; (iv) to dress partial and total lists of both canonical and inflected dictionary entries; (v) to search and list dictionary entries for grammatical categories; (vi) to dress total and partial lists in reverse order.

At first we present some general characteristics of DIGRAMA, in particular those involved in dictionary building. Next we describe briefly the morphological behaviour of the major grammatical classes of Portuguese. Afterwards, we present the solutions we have found to code the linguistic information associated to simple words. Finally, we illustrate how the system generates the inflected forms from canonical entries.

1. General information on the system

1.1. The structure of DIGRAMA

DIGRAMA is being developed in PASCAL and C on a DIGITAL VAX/VMS system. UNIX and MS/DOS versions are also envisaged.

The linguistic information actually dealt with by the system exclusively concerns Portuguese, but DIGRAMA has been conceived to allow the processing of linguistic data from other languages. The fact that the codification of linguistic information is contained in data tables (MIEL - Matrices of specific information on languages) guarantees a large freedom regarding the introduction of new linguistic information, and makes the system largely independent of any particular language information.

Concerning dictionaries, the matrices contain information mainly on grammatical classes of Portuguese and on the inflectional variation of grammatical classes. They also contain the allowed qualifiers for aditional information on words (*eg* morpho-syntactic information to recognize and generate particular verbal and pronominal forms resulting from clitic$-verb$ combinations in Portuguese (see 3.1.1. below)).

1.2. The structure of the system of dictionaries

The system of dictionaries has the following general structure:

- DIGRAS, the dictionary of simple words, is the central element of the system. It contains the simple words in their canonical form: *mesa* (table), *azul* (blue), *pôr* (to put).

We define *simple word* as a string of letters limited by two barriers. A *barrier* is a non alphanumeric character such as the empty space, the hyphen, the dash, and other punctuation signs.

Such a formal definition of simple word led to include in the DIGRAS lexical units such as *azul* (bleu), *mesa* (table), *pôr* (to put), easily recognized as Portuguese words, as well as forms such as *big*, *bang*, *end*, *mouxe*, *sine*, *week*, etc., that are not independent words in Portuguese; nevertheless they have been integrated into the DIGRAS because they are constituents of compound lexical units: *teoria do big bang* (big bang theory), *week-end*, *sine die*, *a trouxe-mouxe* (higgledy-piggledy), *sine qua non*. For the same reasons, some prefixes have also been introduced into the dictionary: *anti-* (*anti-social*), *mini-* (*mini-saia* (mini-skirt)), *multi-* (*multi-racial*), etc.. The contractions of prepositions with determiners, pronouns, adverbs, etc. (eg. *pelo* = *por* + *o* (by + the)); *deste* = *de* + *este* (of + this); *daqui* = *de* + *aqui* (from + here)), very frequent in Portuguese, have also been considered as simple words. The relations between these contracted forms and their constituent categories are established in a specific matrix.

The simple words are introduced into DIGRAS together with alphanumeric codes that allow to integrate them into their word-class(es), and to give them a pattern of morphological variation. This information has been systematically coded, and constitutes the basis for a program of automatic generation of the inflected forms associated to each dictionary entry.

- DIGRAF is made up of all the inflected forms from the words of DIGRAS: *azul*, *azuis* (bleu, sing, pl), *mesa*, *mesas* (table, tables), *pôr*, *ponho*, *pões*,... (to put, I put, you put,...).

- DIGRAC is the dictionary of compound words. A *compound word* is a lexical unit constituted by more than one simple word and any allowed barriers. The adjectives *azul marinho* (navy blue), *azul celeste* (sky-blue), the nouns *mesa redonda* (round table), *pôr-do-sol* (sunset), the adverbs *de repente* (suddenly), *de cor* (by heart), the prepositions *apesar de* (in spite of), *acerca de* (about) are examples of such compound lexical units.

The inflected forms of the compound words constitute DIGRACF. The linguistic description of compound words as well as the programs to process them are being worked out.

We also envisage the construction of dictionaries corresponding to the phonetisation of entries from both DIGRAS and DIGRAC.

2. The dictionary DIGRAS

2.1. The lexical entries of DIGRAS

As mentioned above, DIGRAS is the dictionary that contains simple words. The output structure of the lexical entries is constituted by two elements:

$< word > . < morphological description >$

where *word* represents the *canonical form* of a simple lexical unit, for instance:

masculine, *singular* for nouns and adjectives:
gato (cat), *generoso* (generous);

feminine, *singular* for exclusively feminine nouns and adjectives:
mesa (table), *grávida* (pregnant);

infinitive for verbs:
lavar (to wash), *comer* (to eat);

and *morphological description* corresponds to an alphanumeric code containing informations on the grammatical category of words as well as on their inflectional pattern. The entries: *gato*, *generoso*, *mesa*, *lavar* have the following format:

gato.N001A25
generoso.Adj001S1
mesa.N301A2
lavar.V1F01T

The symbols *N*, *Adj*, *V* indicate that the word is respectively a noun, an adjective and a verb. The remaining characters refer to the pattern of morphological variation of the word: the code *001* corresponds to nouns and adjectives which present the variation *-o*, *-a*, respectively for masculine and feminine forms, and end in *-s* in the plural; the inflectional pattern *301* is that of exclusively feminine nouns and adjectives; *A25*, *A2*, it S1 contain information on diminutive, augmentative and superlative suffixes (see 3.1. and 4.1. below); *V1F01T* indicates that *lavar* is a verb belonging to the first conjugation (*ie* its infinitive form ends in *-ar*), that it has the inflectional pattern *F01*, and can be constructed with all personal pronouns (see below).

2.1.1. Cases of homography

There are numerous lexical entries that are formally identical, but have different syntactic and/or semantic values: they are homographs. At the morphological level it is not possible to distinguish adequately between different cases of homography. At that level of description, word combinations are not analysed, and only the morphological properties of the simple lexical units (grammatical class, inflectional variation) are taken into account. In Portuguese the amount of homographs is considerably high, and they have been assembled under the same entry. Consequently, all the entries of DIGRAS are formally (orthographically) different. Nevertheless the different types of homography have been described in a different way.

Case 1. Homographs belonging to the same grammatical category and to the same inflectional pattern:

A lexical item such as *vela* (candle; sail; spark plug), in spite of its different meanings, has the same morphological behaviour, and constitutes a single entry in DIGRAS:

vela.N301A2

In the same way, homograph verbs like *decorar*: *decorar uma casa* (to decorate a house), *decorar uma lição* (to learn a lesson by heart) are assembled under the same entry:

decorar.V1F01T

Case 2. Homographs that belong to the same grammatical class but have different inflectional and/or morphological behaviour:

This is the case of the noun *macaco* (monkey; jack), with entry:

macaco.N002A4.N200

The code *N002A4* corresponds to the noun that has a masculine and a feminine form for both singular and plural (the animal); *N200* represents the exclusively masculine homograph (the apparatus for lifting a car).

Case 3. Homographs that belong to different grammatical classes

A very frequent case of homography in Portuguese occurs with words that can be used both as nouns and as adjectives. A lexical item such as *louro*:

(1) *O louro disse "olá"* (the blond said "hello")
(2) *O menino louro disse "olá"* (the blond child said "hello")

may be found, as these examples show, in a nominal, (1), or adjectival position, (2). Therefore the word has been classified as noun and as adjective, and the corresponding inflectional information has been coded as follows:

louro.N023A9.N223.Adj023S9

However, as a noun, *louro* is still ambiguous since it can refer to a human, in which case it inflects in number and gender (code *N023*), or to an animal (a parrot), presenting then masculine forms only (code *N223*). In fact, example (1) above allows both interpretations: *the blond said "hello"*; *the parrot said "hello"*. As mentioned above, this kind of ambiguity cannot be adequately solved at the morphological level of description.

The lexical entry *capital* (capital) is another example of that type of homography:

capital.N211A11.N311.Adj111

It may either be a noun or an adjective. The code *N211* corresponds to the masculine noun *o capital*, *os capitais* (the money); *N311* represents the exclusively feminine homograph *a capital*, *as capitais* (town, towns). The code *Adj111* means that *capital* may also be an adjective: *um pecado capital* (a capital sin), *uma pena capital* (a capital punishment) belonging to the inflectional class *111*.

Before giving further informations on the codification of linguistic information, we present the main morphological variations of Portuguese grammatical categories.

3. The inflectional variation of major grammatical classes in Portuguese

As in other Romance languages, grammatical categories in Portuguese may be morphologically variable and invariable.

Variable categories: **N**(ouns), **Adj**(ectives), **V**(erbs), some **Adv**(erbs), some **Pron**(nouns) and some **Det**(erminers);

Invariable categories: **Prep**(ositions), **Conj**(unctions), **Inter**(jections), the majority of **Adv**(erbs), some **Pron**(ouns) and some **Det**(erminers).

3.1. Morphological variation of nouns and adjectives

Besides the variation in gender (masculine, feminine) and number (singular, plural):

N: Ms, Fs, Mp, Fp:
gato, *gata*, *gatos*, *gatas* (cat)

Adj: Ms, Fs, Mp, Fp:
generoso, *generosa*, *generosos*, *generosas* (generous)

nouns and adjectives often present diminutive and augmentative suffixes that express either quantity (big, small) or quality (affection, irony). The most frequent diminutives are: -*inho*, -*ito*, -*zinho*, -*zito* (coded: **D**_inho, **D**_ito, **D**_zinho, **D**_zito); the more general augmentative suffix is -*ão* (coded: **A**_ão). Thus, a noun like *gato* (cat) is associated to the following forms:

Ms, Fs, Mp, Fp: *gato*, *gata*, *gatos*, *gatas* (cat)
D_inho: *gatinho*, *gatinha*, *gatinhos*, *gatinhas* (small cat; little cat)
D_ito: *gatito*, *gatita*, *gatitos*, *gatitas* (small cat; little cat)
A_ão: *gatarrão*, *gatarrona*, *gatarrões*, *gatarronas* (big cat; great cat)

Regarding adjectives, the superlatif degree (that expresses the highest degree) of gradable adjectives is also expressed morphologically by inflected forms in -*íssimo*, -*ílimo*, -*érrimo*, etc. An adjective such as *gordo* (fat) may take inflectional forms that include:

Ms, Fs, Mp, Fp: *gordo*, *gorda*, *gordos*, *gordas* (fat)
D_inho: *gordinho*, *gordinha*, *gordinhos*, *gordinhas* (rather fat)
D_ito: *gordito*, *gordita*, *gorditos*, *gorditas* (rather fat)
S_íssimo: *gordíssimo*, *gordíssima*, *gordíssimos*, *gordíssimas* (extremely fat)

As mentioned before, these suffixes are used to achieve intensifying effects similar to those that could also be expressed by syntactic means (*eg*: by using periphrastic forms, adjectives, adverbs, and certain noun phrases, for example: *ele é bastante gordo* (he is rather fat), *ele é um bocado gordo* (he is a little bit fat)).

In general, as the preceding examples show, the inflectional variation of nouns and adjectives only affects their endings: the marks of gender and number are carried out by the diminutive, augmentative, superlative suffixes added directly to the stem of the word. Thus, the feminine and the plural forms of a lexical item such as **N**, **Adj**: *louro* (blond):

Ms, Fs, Mp, Fp: *lour* (o, a, os, as)
D_inho: *lourinh* (o, a, os, as)
D_ito: *lourit* (o, a, os, as)
S_íssimo: *louríssim* (o, a, os, as)

appear only in the ending of suffixes, directely added to the stem (*lour*-) of the word.

However, with the diminutive suffixes -*zinho* and -*zito*, we find a different situation. These suffixes are not directly added to a stem but to an inflected word. Sometimes the forms -*zinho* and -*zito* can alternate with -*inho* and -*ito* but, in general, nouns and adjectives stressed on the last syllabe, as well as the monossyllabic ones, only accept the diminutives -*zinho* and -*zito*. This is the case of **N**, **Adj**: *anão* (dwarf) and **Adj**: *nu* (naked):

Ms, Fs, Mp, Fp: *anão*, *anã*, *anões*, it anãs
D_zinho: *anãozinho*, *anãzinha*, *anõezinhos*, *anãzinhas*
D_zito: *anãozito*, *anãzita*, *anõezitos*, *anãzitas*
Ms, Fs, Mp, Fp: *nu*, *nua*, *nus*, *nuas*
D_zinho: *nuzinho*, *nuazinha*, *nuzinhos*, *nuazinhas*

As these examples show, the affixes of gender and number, *eg*: an**õe**zinho**s** (plural masculine) and not **anãozinhos*, nu**a**zinh**a** (singular feminine) and not **nuzinha*, appear twice: inside and in the ending of the word.

If the word that takes the suffixe has a regular plural form, *ie* adjunction of -*s* to the singular, for example the masculine noun *café* (coffee):

Ms, Mp: *café*, *cafés*
D_zinho: *cafezinho*, *cafezinhos*

the morpheme of the plural, -*s*, is assimilated to the sibilant, -*z*-, of the suffix, and the mark of the plural is only seen in the ending.

3.2. Inflection of adverbs

For a small number of adverbs, the inflected forms used for diminutives and superlatives are the same as those for adjectives:

Adv: *cedo* (early)
D_inho: *cedinho* (very/rather early)
D_ito: *cedito* (very/rather early)
S_íssimo: *cedíssimo* (very, very early)

The majority of adverbs, however, are invariable. This is the case of adverbs ending in -*mente* (-ly). Those adverbs have different syntactic and semantic values, and are mainly formed by adding the derivational suffix -*mente* to the feminine base form of an adjective:

dura > *duramente* (hard (feminine) > hardly)
rara > *raramente* (rare (feminine) > rarely)

Some adverbs in -*mente* are also derived from the feminine superlative form of adjectives:

duríssima > *durissimamente* (very, very hardly)
raríssima > *rarissimamente* (very, very rarely)

This is also a means to achieve an intensifying effect. For a few adverbs ending in -*mente*, the superlative form of the adjective is their unique derivational base. The adverb *pessimamente* (most badly) is an example of that situation. It is derived from the adjective *mau* (bad), which is irregulary inflected:

Ms, Fs, Mp, Fp: *mau*, *má*,*maus*,*más* (bad)
Comparative: Ms, Fs: *pior*, Mp, Fp: *piores* (worse)
Superlative: *péssimo*, *péssima*, *péssimos*, *péssimas* (worst)

The adverb is formed from the feminine superlative *péssima*, and there is no adverb **mamente* (badly) derived from the base *má*.

The addition of the suffixes, -*inho*, -*zinho*, -*mente*, etc., to a base involves an accentual change, and, consequently, the lost of the accent, if any. For example:

café > *cafezinho*
péssima > *pessimamente*

The noun *café* is stressed on the last syllable, the derived form *cafezinho* on the penultimate; the superlative adjective (*péssima*) is stressed on the antepenultimate syllable, the adverb on the penultimate.

On the other hand, the superlative forms of adjectives and adverbs, base + -*íssimo*, are always stressed on the antepenultimate syllable:

gordo, *gordíssimo*
cedo, *cedíssimo*

an accentual position orthographically marked.

The addition of diminutive and superlative suffixes can also change the orthographic form of the stem:

N, **Adj**: *amig* (o, a, os, as) (friend)

D_inho: *amigu* (inho, inha, inhos, inhas)
D_ito: *amigu* (ito, ita, itos, itas)
Adj Super: *amic* (íssimo, íssima, íssimos, íssimas)

For phonetic reasons, the addition of *-inho* and *-ito* to the nominal and adjectival stem *amig-* leads to the appearance of *-u-* (which inhibits the palatalization of the *-g-*); the superlative of the adjective *amigo* is formed from an erudite base directly related to the latin form *amicus*.

Thus, the possibility for a word to accept such or such derivational suffix has been codified, in order to obtain the correct generation of those inflected forms.

The formal changes induced by such suffixes led to an augmentation of inflectional patterns. The words *amigo* and *gordo* have an identical inflection for gender and number:

amig(o, a, os, as)
gord(o, a, os, as)

but behave differently in the presence of diminutive and superlative suffixes:

amigu(inho, inha, inhos, inhas)
gord(inho, inha, inhos, inhas)
amic(íssimo, íssima, íssimos, íssimas)
gord(íssimo, íssima, íssimos, íssimas)

The stem *amig-* presents formal variants; the stem *gord-* remains unchanged. To *amigo* corresponds the inflectional code *Adj004S3*, to *gordo* the code *Adj001S1*.

In total, the nouns and adjectives have been devided into 135 and 151 productive inflectional classes, respectively.

3.3. Morphological variation of verbs

According to the ending of their infinitive forms, the verbs belong to four conjugations:

-ar: *lavar* (to wash)
-er: *comer* (to eat)
-ir: *partir* (to leave)
-or: *pôr* (to put)

Inside each conjugation, the verbs can inflect in *Mood* (indicative, subjunctive), *Tense* (present, past, future, etc.), *Person* and *Number*. They also have invariable forms: *infinitive*, *gerundive*, and *past participle* (which is used in compound tenses of the verbs).

Simple tenses of the indicative mood are:

Present: lav(o, as, a, amos, ais, am) = 6 forms
Imperfect: lav(ava, avas, ava, ávamos, ávais, avam) = 6 forms
Perfect: lav(ei, aste, ou, ámos, astes, aram) = 6 forms
Pluperfect: lav(ara, aras, ara, áramos, áreis, aram) = 6 forms
Future: lav(arei, arás, á, aremos, areis, arão) = 6 forms
Conditional: lav(aria, arias, aria, aríamos, aríeis, ariam) = 6 forms

Simple tenses of the subjunctive mood are:

Present: lav(e, es, e, emos, eis, em) = 6 forms
Imperfect: lav(asse, asses. asse, ássemos, asseis, assem) = 6 forms
Future: lav(ar, ares, ar, armos, ardes, arem) = 6 forms

The imperative has some specific forms according to its occurrence in positive or negative sentences. The *positive* forms are:

Positive imperative: lav(a, e, emos, ai, em) = 5 forms

These forms are partially identical to those appearing in negative sentences:

Negative Imperative: lav(es, e, emos, eis, em) = 5 forms

The positive or negative forms of the imperative correspond to: two forms for second person singular: *lava* (positive), *laves* (negative) when there is an informal relationship of the speaker and the hearer; *lave* (formal relationship in both positive and negative sentences); one form for the first person plural: *lavemos*; three forms for second person plural: *lavai*, *laveis*(two old-fashioned forms used in positive and negative sentences respectively), *lavem* (standard form for both formal and informal relationship, positive and negative sentences). In the other verb tenses, the formal relationship of the speaker and the hearer is expressed by the third persons (singular or plural).

There are two types of infinitive forms in Portuguese: an invariable form: *lavar*, and an inflected (person and number) one:

Inflected infinitive: lav(ar, ares, ar, armos, ardes, arem) = 6 forms

This means that, in Portuguese, a non-impersonal and non-defective verb, such as *lavar*, gives rise to 73 simple forms, including some homographs.

Compound tenses of the verbs are formed by the auxiliary *ter* (to have) and the past participle of lexical verbs (*Vpp*). Since compound tenses consist of a string of two words: *ter* + *Vpp*, they are not included in the dictionary of simple words (DIGRAS). Let us however make some remarkes on the morphological behaviour of past participle in Portuguese. Contrarily to what is observed in some Romance languages (*eg* in French), in compound tenses of Portuguese no person and number concord is found between the subject or the direct object and the past participle:

*Ele tem comid***o** *chocolates todos os dias*
(He has been eating chocolates every day)

*Ela tem comid***o** *chocolates todos os dias*
(She has been eating chocolates every day)

*Esses rapazes tenho-os vist***o** *regularmente*
(Those boys I have been seeing them regularly)

*Essas raparigas tenho-as vist***o** *regularmente*
(Those girls I have been seeing them regularly

These examples show that there is no concord between the participle *comido* (lit. eaten) and the subject of the verb, *ele*, *ela* (he, she); nor there is any agreement between the participle *visto* (lit. seen) and the direct object *esses rapazes* (those boys), *essas raparigas* (those girls).

Therefore, we have considered that the past participle is an invariable form, and have distinguished it from other variable participle-like forms appearing in:

a) passive sentences:

*Esses rapazes têm sido vist***os** *regularmente*
(Those boys have been seen (plur. masc.) regularly)

*Essas raparigas têm sido vist***as** *regularmente*
(Those girls have been seen (plur. fem.) regularly)

b) passive-like constructions:

*Esses rapazes estão bronzead***os**
(Those boys are bronzed (plur. masc.))

*Essas raparigas estão bronzead***as**
(Those girls are bronzed (plur. fem.))

c) adjectival position:

*Um rapaz bronzead***o**
(A bronzed boy)

*Uma rapariga bronzead***a**
(A bronzed girl)

3.3.1. Formal changes induced by pronouns

The verb forms, ending in *-s*, *-r*, *-z*, can undergo some modifications caused by the presence of the accusative, dative and reflexive forms of personal pronouns (*clitics*), which, in turn, can simultaneously undergo formal alterations. Therefore, the eventual co-occurrence of such pronominal forms with the verbs had to be taken into account at the morphological level of description. At first we give a brief description of the position of clitics in Portuguese, later we point out the main formal changes found in such verb−*pronoun* combinations.

In european Portuguese, the accusative, dative and reflexive forms of personal pronouns are, in simple declarative sentences, attached to the verbs by means of an hyfen, contrarily to what is observed in brasilian Portuguese, where such clitics come in a pre-verbal position:

European Portuguese: *Ele lava*-**o** (He washes it/him)
Brasilian Portuguese: *Ele* **o** *lava* (He it/him washes)

The pre-verbal position is also observed in european Portuguese in the following sentence types:

Negative sentences: *Ele não* **o** *lava* (He does not it/him wash)
Subordinate clauses: *Que ele* **o** *lave* (That he it/him wash)
Wh-questions: *Quem* **o** *lava ?* (Who it/him washes?)
Wh-exclamations: *Como ele* **o** *lava*! (How he it/him washes!)

The clitics can also be attracted to a pre-verbal position by some pronouns, adverbs and quantifiers:

Alguém **o** *lava* (Somebody it/him washes)
Ele já **o** *lava* (He already it/him washes)

In the future tense as well as in the conditional, the clitics are inserted in the verb form:

Future: *Ele lavá*-**lo**-*á* (He will wash it/him)
Conditional: *Ele lavá*-**lo**-*ia* (He would wash it/him)

-lo- is a positional variant of **o**, and it separates the forms *lavará* and *lavaria* (third person singular of future and conditional) into their etymological constituents: the infinitive of a lexical verb (*lavar*) plus the endings of the present and imperfect of the auxiliary verb *haver* (3th, sing. *há*, *havia*). As the examples above show, in the future and in the conditional both the verb and the pronoun undergo morphological changes.

The main morphological alterations due to pronouns are found when there is a combination of a verb form ending in -*s*, -*r*, -*z*, and a pronoun constituted by a vowel (some accusative forms). Such eventuality will be illustrated by conjugating the indicative present of the regular verb *lavar*, alone, and combined with an accusative clitic, *o* (third person masculine singular, for both human and non-human referents):

Eu lavo		*Eu lavo-o*	(I wash it/him)
Tu lavas	–>	*Tu lava-lo*	(You wash it/him)
Ele lava		*Ele lava-o*	(He washes it/him)
Nós lavamos	–>	*Nós lavamo-lo*	(We wash it/him)
Vós lavais	–>	*Vós lavai-lo*	(You wash it/him)
Eles lavam	–>	*Eles lavam-no*	(They wash it/him)

It can be observed that the adjunction of a vocalic clitic to a verb form ending in -*s* induces formal changes both in the verb (lost of the *s*) and in the pronoun (*o* –> *lo*); in turn, the verb forms ending in a nasal sound (*eg*/m/) transmit the nasalization to the pronouns (*o* –> *no*).

The combination of the verbs with non-vocalic pronouns also lead to a few formal alterations of the verb forms, such as the ones illustrated by:

Nós lavamos + nos –> Nós lavamo-nos (We wash ourselves)

Thus, even though the complete description of the possibilities of combination of a verb with a pronoun is a question to be dealt with at the syntactic level, given the morphological changes induced by such combinations in Portuguese, a first treatment of that subject had to be made at the morphological level.

3.3.2. Morphological classification of verbs

The morphological classification of verbs has been established according to two major criteria: (i) their formal pattern of variation (*ie* the verbs have been included in morphological (orthographic) classes); (ii) the possibility of being or not being followed by the accusative, dative, and reflexive forms of personal pronouns.

The adoption of criterion (ii) already allows to attach to the verb entries a few syntactic informations, that will be later refined and completed at the syntactic level. At the present stage of description, the verbs have been subdivided into four groups: those verbs that accept both accusative and dative pronouns; those that can only appear with accusative forms (they only take a direct object); those that can be combined with dative forms (they take an indirect object, usually preceded by the preposition *a* (to)); finally, those verbs that are exclusively intransitive.

Thus, a verb entry of the DIGRAS can have the following format:

*Word.*V1F33T

which means that *word* is a verb (*V*) belonging to the first conjugation (1), that it has the inflectional pattern *33*, and can accept all the clitic pronouns (*T*). The entries corresponding to the verbs *lavar* (to wash), *amar* (to love), *telefonar* (to telephone), and *jejuar* (to fast) have the form:

lavar.V1F01T
amar.V1F01A
telefonar.V1F01D
jejuar.V1F01N

All these verbs belong to the first conjugation (they end in -*ar*), and take the inflectional pattern 01 (regular verbs of the first conjugation). They have been, however, included into four different sub-classes, according to the type of clitic that they combine with; *jejuar* is an intrasitive verb, it does not take any pronoun.

We have established 83 morphological classes where have been integrated the 6,000 verb entries formalized so far.

4. The codification of linguistic informations

As said before, the linguistic information that can be assigned to isolated words is mainly morphological: grammatical category of the words, inflectional patterns of the different grammatical categories. The introduction of a simple lexical unit into DIGRAS is accompanied by an alphanumeric code that associates a given word base form to a grammatical class and to an inflectional pattern. Naturally that pressuposes that the inflection patterns of the different grammatical classes have previously been introduced into the system.

4.1. Codification of nouns and adjectives

In order to illustrate the system operation, let us consider a simple example: the introduction of a noun such us *gato* (cat). This entry is associated to the following word forms:

gat*o*, gat*a*, gat*os*, gat*as*
gat*inho*, gat*inha*, gat*inhos*, gat*inhas*
gat*ito*, gat*ita*, gat*itos*, gat*itas*
gat*arrão*, gat*arrona*, gat*arrões*, gat*arronas*

The first line represents what may be called the basic forms of the word, the remaining lines correspond to two diminutives and one augmentative respectively. The example shows that all the inflected forms consist of a commun segment: *gat-*, followed by a variable string of characters. Hence, to generate the set of words above, the following structure is used:

(o, a, os, as)
(inho, inha, inhos, inhas)
(ito, ita, itos itas)
(arrão, arrona, arrões, arronas)

the first line corresponds to the inflectional rule for gender and number of the stem *gat-*; the others represent the diminutive and augmentative suffixes accepted by the stem. This information is introduced by a command of the form:

DEF FLEX N001A25
(base (@o, a, os, as),
D_inho (inho, inha, inhos, inhas),
D_ito (ito, ita, itos, itas),
A_ão (arrão, arrona, arrões, arronas))

The sign ”@” indicates that the form corresponds to the canonical form of the lexical unit (in this example the singular masculine).

Therefore, the introduction of the noun *gato* is achieved by associating its canonical form to the following inflectional rules:

DEF N gat$< o >$ FLEX N001A25

The canonical form *gat*$< o >$ is constituted by an invariant string, and by a variable segment, enclosed between $< ... >$. The latter corresponds to the part of the word that, when substituted for the elements specified in its inflectional pattern, allows the generation of all the inflected forms associated to that word.

Some words may have more than one equivalent form. This is the case for the noun *aldeão* (villager), which has three possible different forms for plural masculine: *aldeões*, *aldeãos*, *aldeães*. In such cases, the position corresponding to the plural masculine is filled with the equivalent forms, linked up by the "=" sign:

DEF FLEX N041
(base (@ão, ã, ões=ãos=ães, ãs))

The method that has been used to link the invariable form of a word to a given inflectional pattern allows for the substitution of both the endings of the words, and/or any segment inside the words. In fact, there are numerous lexical items that, besides the eventual inflection in number and gender, may have positional variants corresponding to a formal change inside de word. For such cases, *eg*: *to***u***ro* - *to***i***ro* (bull), the corresponding variational pattern is:

DEF FLEX N223 N
(base (@<u><o>=<i><o>,,<u><os>=<i><os>,))

These entries are introduced into the dictionary in the following way:

DEF N to<u>r<o> FLEX N223

The inflectional properties of adjectives are identical to those of the nouns. As mentioned before, they may vary in gender and number; they often accept diminutive and augmentative sufixes, and, when gradable, they may form the superlative by morphological means. The codification of their morphological behaviour do not differ much from that of nouns. Thus to an adjective such as *louro* (blond), that, like the noun *touro*, has an equivalent form *loiro*, corresponds the variational pattern:

DEF FLEX ADJ023S9 ADJ
(base (@ <u><o>=<i><o>,<u><a>=<i><a>,
<u><os>=<i><os>,<u><as>=<i><as>),
D_inho (<u><inho>=<i><inho>,<u><inha>=<i><inha>,
<u><inhos>=<i><inhos>,<u><inhas>=<i><inhas>),
D_ito (<u><ito>=<i><ito>,<u><ita>=<i><ita>,
<u><itos>=<i><itos>,<u><itas>=<i><itas>),
S_issimo (<u><íssimo>=<i><íssimo>,<u><íssima>=<i><íssima>,
<u><íssimos>=<i><íssimos>,<u><íssimas>=<i><íssimas>))

The entry *louro* is introduced into the dictionary by:

DEF Adj lo<u>r<o> FLEX Adj023S9

4.2. Codification of verbs

Regarding verbs, the information that has been coded concerns, on the one hand, their inflection in person, number, tense, and mood; on the other hand, the formal changes induced by the clitics.

As said before, the combination of a verb form with a clitic often leads to a formal change of the verb, of the clitic, or of both verb and clitic. The possibility for a given verb form of being or not being followed by a clitic has been codified in the following way:

V (..., $\mathrm{s}c_1 ... c_n$: V$\mathrm{fle}x_1 = c_{n+1} ... c_m$: V$\mathrm{fle}x_2, ...$)

$sc_1 ... c_n$ indicates that the verb form *Vflex*$_1$ is not followed by an hyphen (code s) or that it is followed by an hyphen and a clitic that does not induces formal changes of the verb; *Vflex*$_2$ corresponds to a deformed equivalent of *Vflex*$_1$ resulting from the combination of *Vflex*$_1$ and a clitic of the series $c_{n+1} ... c_m$.

Thus a verb such as *lavar* (to wash) – a regular verb of the first conjugation that combines with accusative, dative and reflexive clitics – is introduced into the dictionary by associating to its infinitive form the inflectional code:

```
DEF V lav<ar> FLEX V1F01T
```

The inflectional pattern corresponding to *V1F01T* is:

```
DEF FLEX V1F01T V
(ip     (SAQVO:o,SAQV:as= L:a,SAQVO:a,
        SAV:amos= QL:amo,SAQV:ais= L:ai,SAQVN:am),
ipi     (SAQVO:ava,SAQV:avas= L:ava,SAQVO:ava,
        SAV:ávamos= QL:ávamo, SAQV:áveis= L:ávei,SAQVN:avam),
ipp     (SAQVO:ei,SAQVO:aste,SAQVO:ou,
        SAV:ámos= QL:ámo,SQVA:astes= L:aste,SAQVN:aram),
imqp    (SAQVO:ara,SAQV:aras= L:ara,SAQVO:ara,
        SAV:áramos= QL:áramo,SAQV:áreis= L:árei,SAQVN:aram),
if      (S:arei,S:arás,S:ará,S:aremos,S:areis,S:arão),
ic      (S:aria,S:arias,S:aria,S:aríamos,S:aríeis,S:ariam),
ifcr    (AQV:ar= L:á),
cp      (S:e,S:es,S:e,S:emos,S:eis,S:em),
cpi     (S:asse,S:asses,S:asse,S:ássemos,S:asseis,S:assem),
cf      (S:ar,S:ares,S:ar,S:armos,S:ardes,S:arem),
imp     (SAQVOZ:a= SY:es,SYZ:e= AQVOZ:e,
        SYZ:emos= AVZ:emos= QLZ:emo,
        SAQVOZ:ai= SY:eis,SYZ:em= AQVNZ:em),
iip     (@SAQV:ar= L:á),
ifx     (SAQV:ar= L:á,SAQV:ares= L:are=*L:á,SAQV:ar= L:á,
        SA:armos= QL:armo,SAQ:ardes= L:arde,SAQN:arem),
ger     (SAQVO:ando),
pp      (S:ado))
```

Concerning the imperative (*imp*), the codes *Z* and *Y* indicate that the forms are found in positive and/or negative sentences respectively.

5. The dictionary DIGRAF

DIGRAF is automatically generated from the association of a canonical entry to a inflectional code. That dictionary consists then of the canonical forms of DIGRAS and all the morphological variants associated to them. The lists of inflected forms, automatically obtained, have a format as shown in the following lexical family:

amicíssima.Adj004S3:ss_fs	amiga.V1F18R:ip_2's/3s
amiga.N001A6:fs	amigou.V1F18R:ipp_2's/3s
amiga.Adj004S3:fs	amiguinhos.N001A6:dh_mp

The first code indicates the grammatical category and the inflectional pattern of the word; the code next to the colon specifies the inflection of the entry: *ss* stands for superlatives of the form *-íssimo*; *dh* represents the diminutive *-inho*; *m*, *f*, *s*, *p* stand for masculine, feminine, singular and plural respectively; *ip* and *ipp* represent the verb tenses present indicative and past perfect indicative respectively; *2's/3s* correspond to the homograph person forms: second person singular (expressing a formal relationship of the speaker and the hearer, see 3.3.) and third person singular.

6. Using the information of DIGRAS and DIGRAF

The dictionaries DIGRAS and DIGRAF are closely interrelated and, in a sense, constitute a single device. The information contained in both dictionaries enables the system to carry out the following operations on words:

- given a word form, the system recognizes it, shows the grammatical class(es) it belongs to, and specifies the particular inflection that the word presents. A word such as *capital* (capital) will be described as (see 2.1.1. above):

capital.N:ms
capital.N:fs
capital.Adj:mfs

- given a word form, the system identifies it and lists all the inflected forms associated to that word. A noun such as *pato* (duck) will be included into the morphological family:

pato.N

pato.ms;	pata.fs;	patos.mp;	patas.fp
patinho.dh_ms;	patinha.dh_fs;	patinhos.dh_mp;	patinhas.dh_fp
patito.dt_ms;	patita.dt_fs;	patitos.dt_mp;	patitas.dt_fp

- to dress total and partial lists of both canonical and inflected entries (see appendices A, B);

- to search and list dictionary entries for grammatical categories (appendix C)

The reverse dictionaries DIGRASI and DIGRAFI, that present the words in a reverse order, are automatically generated respectively from DIGRAS and DIGRAF. The lists that it is possible to obtain from the reverse dictionaries are of the same type as those obtained from DIGRAS and DIGRAF.

7. Final remarks: some statistics

The dictionary DIGRAS has for the time being (January 1993) about 60,000 lexical entries. As mentioned before all the entries of DIGRAS are formally different. This means that a number of homographs are assembled under the same entry. In terms of lexical categories, 35,000 nouns, 19,000 adjectives, 6,000 verbs, and 1,500 adverbs have been entered.

The inflection patterns established so far to account for the morphological variations of the lexical entries are distributed as follows: 135 for nouns, 151 for adjectives, 83 for verbs, 15 for adverbs.

The dictionary of inflected forms, DIGRAF, contains about 600,000 entries. The inflected forms of nouns and adjectives correspond to the morphological expression of gender and number, as well as of quantification and intensification effects involving diminutive, augmentative, and superlative (for adjectives) suffixes. The inflected forms of the verbs include inflection in person, number, tense, and mood as well as formal changes induced by the clitic pronouns.

If one takes into account that the *Dicionário da Língua Portuguesa* (1991), which are beeing used to collect the data, contains about 80,000 entries, that include a few compound words, the DIGRAS, in its actual stage of development, corresponds to a significative portion of that dictionary.

However the *Dicionário da Língua Portuguesa* is not complete, as it usually happens with dictionaries build for human users. Many words regularly derived from others are often absent (this is the case of many adverbs ending in -*mente*: *rapidamente* (rapidly), *raramente* (rarely), regularly obtained from adjectives, and of adjectives ending in -*vel*: *apagável* (removable), *dedutível* (deductible), associated to transitive verbs (*apagar*, *deduzir*)). Lexical lacunae are frequent. Therefore it is not difficult to foreseen that to reach a reasonable degree of completion DIGRAS should contain not less than 100,000 simple words.

NOTES

We wish to thank Blandine Courtois who has given us so much information from her own experience on electronic dictionary building.

REFERENCES

Almeida Costa, J.; A Sampaio e Melo. 1991. *Dicionário da Língua Portuguesa*, 6^a edição, Porto: Porto Editora.

Courtois, Blandine. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française* 87, Paris: Larousse.

Cunha, Celso; L. Lindley Cintra. 1984. *Nova Gramática do Português Contemporâneo*. Lisboa: Edições João Sá da Costa.

Gross, Maurice. 1988. Methods and Tactics in the Construction of a Lexicon-grammar. In *Linguistics in the Morning Calm* 2, Seoul: Hanshin Publishing Company.

Gross, Maurice. 1989. La constructions de dictionnaires électroniques. *Annales des télécommunications* 44, n^o 1-2, Paris: CNET.

Laporte, Eric. 1990. Le dictionnaire phonémique DELAP. *Langue Française* 87, Paris: Larousse.

Marques Ranchhod, Elisabete. 1989. Lexique-grammaire du portugais: Prédicats nominaux supportés par **estar**. *Linguisticae Investigationes* XIII: 2, Amsterdam: John Benjamins B.V.

Marques Ranchhod, Elisabete. 1990. *Sintaxe dos predicados nominais com* **estar**. Lisboa: Instituto Nacional de Investigação Científica.

Marques Ranchhod, Elisabete. 1991. Frozen Adverbs - Comparative Forms **como C** in Portuguese. *Linguisticae Investigationes* XV: 1, Amsterdam: John Benjamins B.V.

Marques Ranchhod, Elisabete; Eleutério, Samuel. (forthcoming). As novas tecnologias e o estudo do português. In *III Encontro da Associação das Universidades de Língua Portuguesa*, Lisboa: AULP.

Sá Nogueira, Rodrigo. 1978. *Dicionário de verbos portugueses conjugados*. Lisboa: Livraria Clássica Editora.

Silberztein, Max. 1990. Le dictionnaire électronique des noms composés. *Langue Française* 87, Paris: Larousse.

Appendix A

Examples from DIGRAS

denotação.N308
denotado.N200.Adj001
denotador.Adj005
denotar.V1F01A
denotativo.Adj001
densamente.Adv1
densidade.N300
densifloro.Adj001
densifoliado.Adj001
densimetria.N300
densímetro.N200
densissimamente.Adv1
denso.Adj001S1
dentada.N300A2
dentado.Adj001
dentadura.N300A12
dental.Adj111
dentalgia.N300
dentálio.N200
dentão.N208
dentar.V1F01T
dentário.Adj001
dente.N200A2
denteação.N308
denteado.Adj001
dentear.V1F07T
dentebrum.N210
dentebrura.N300
denteira.N300
dentel.N214
dentelária.N300
dentelete.N200
dentelha.N300
dentelo.N200
dentição.N308
denticórneo.Adj001
denticulado.Adj001
denticular.Adj105
dentículo.N200
dentiforme.Adj101
dentifrício.N200
dentífrico.N200.Adj001
dentígero.Adj001
dentilabial.Adj111
dentilária.N300
dentilha.N300
dentilhão.N208
dentina.N300
dentípode.Adj101
dentirrostro.Adj001
dentista.N101
dentola.N300
dentolabial.Adj111
dentolas.N116
dentolingual.Adj111
dentoneira.N300
dentre.Prep&Prep
dentro.Adv1
dentuça.N300
dentuças.N116
dentudo.Adj001
denudação.N308
denudar.V1F01T
denúncia.N300
denunciação.N308
denunciado.Adj001
denunciador.N005.Adj005
denunciante.N101.Adj101
denunciar.V1F01T
denunciativo.Adj001
denunciatório.Adj001
denunciável.Adj111
denutrição.N308
denutrido.Adj001
denutriente.Adj101
denutrir.V3F01A
deodáctilo.Adj001
deontologia.N300
deontológico.Adj001
deontologismo.N291
deoperculado.Adj001
deparador.Adj005
departamental.Adj111
departamento.N200
departição.N308
departimento.N200
departir.V3F01T
depascente.Adj101

Appendix B

Examples from DIGRAF

densa.Adj001S1:fs
densamente.Adv1
densas.Adj001S1:fp
densidade.N300:fs
densidades.N300:fp
densiflora.Adj001:fs
densifloras.Adj001:fp
densifloro.Adj001:ms
densifloros.Adj001:mp
densifoliada.Adj001:fs
densifoliadas.Adj001:fp
densifoliado.Adj001:ms
densifoliados.Adj001:mp
densimetria.N300:fs
densimetrias.N300:fp
densímetro.N200:ms
densímetros.N200:mp
densíssima.Adj001S1:ss_fs
densissimamente.Adv1
densíssimas.Adj001S1:ss_fp
densíssimo.Adj001S1:ss_ms
densíssimos.Adj001S1:ss_mp
denso.Adj001S1:ms
densos.Adj001S1:mp
denta.V1F01T:ip_2s(l)/2's/3s:imp_2s(z)
dentá.V1F01T:ifcr(l):iip(l):ifx_1s(l)/2s(*l)/2's(l)/3s(l)
dentada.N300A2:fs
dentada.Adj001:fs
dentadas.N300A2:fp
dentadas.Adj001:fp
dentadinha.N300A2:dh_fs
dentadinhas.N300A2:dh_fp
dentadita.N300A2:dt_fs
dentaditas.N300A2:dt_fp
dentado.Adj001:ms
dentado.V1F01T:pp
dentados.Adj001:mp
dentadura.N300A12:fs
dentaduras.N300A12:fp
dentadurazinha.N300A12:dh_fs
dentadurazinhas.N300A12:dh_fp
dentadurazita.N300A12:dt_fs
dentadurazitas.N300A12:dt_fp
dentai.V1F01T:ip_2p(l):imp_2p(z)
dentais.Adj111:mfp
dentais.V1F01T:ip_2p
dental.Adj111:mfs
dentalgia.N300:fs
dentalgias.N300:fp
dentálio.N200:ms
dentálios.N200:mp
dentam.V1F01T:ip_2'p/3p
dentamo.V1F01T:ip_1p(lq)
dentámo.V1F01T:ipp_1p(lq)
dentamos.V1F01T:ip_1p
dentámos.V1F01T:ipp_1p
dentando.V1F01T:ger
dentão.N208:ms
dentar.V1F01T:ifcr:cf_1s/2's/3s:iip:ifx_1s/2's/3s
dentara.V1F01T:imqp_1s/2s(l)/2's/3s
dentará.V1F01T:if_2's/3s
dentaram.V1F01T:ipp_2'p/3p:imqp_2'p/3p
dentáramo.V1F01T:imqp_1p(lq)
dentáramos.V1F01T:imqp_1p
dentarão.V1F01T:if_2'p/3p
dentaras.V1F01T:imqp_2s
dentarás.V1F01T:if_2s
dentarde.V1F01T:ifx_2p(l)
dentardes.V1F01T:cf_2p:ifx_2p
dentare.V1F01T:ifx_2s(l)
dentarei.V1F01T:if_1s
dentárei.V1F01T:imqp_2p(l)
dentareis.V1F01T:if_2p
dentáreis.V1F01T:imqp_2p
dentarem.V1F01T:cf_2'p/3p:ifx_2'p/3p
dentaremos.V1F01T:if_1p
dentares.V1F01T:cf_2s:ifx_2s
dentaria.V1F01T:ic_1s/2's/3s
dentária.Adj001:fs
dentariam.V1F01T:ic_2'p/3p
dentaríamos.V1F01T:ic_1p
dentarias.V1F01T:ic_2s
dentárias.Adj001:fp
dentaríeis.V1F01T:ic_2p
dentário.Adj001:ms
dentários.Adj001:mp
dentarmo.V1F01T:ifx_1p(lq)
dentarmos.V1F01T:cf_1p:ifx_1p

Appendix C

Lists of Nouns and Adjectives

denotação.N308
denotado.N200
densidade.N300
densimetria.N300
densímetro.N200
dentada.N300A2
dentadura.N300A12
dentalgia.N300
dentálio.N200
dentão.N208
dente.N200A2
denteação.N308
dentebrum.N210
dentebrura.N300
denteira.N300
dentel.N214
dentelária.N300
dentelete.N200
dentelha.N300
dentelo.N200
dentição.N308
dentículo.N200
dentifrício.N200
dentífrico.N200
dentilária.N300
dentilha.N300
dentilhão.N208
dentina.N300
dentista.N101
dentola.N300
dentolas.N116
dentoneira.N300
dentuça.N300
dentuças.N116
denudação.N308
denúncia.N300
denunciação.N308
denunciador.N005
denunciante.N101
denutrição.N308
deontologia.N300
deontologismo.N291
departamento.N200
departição.N308

denotado.Adj001
denotador.Adj005
denotativo.Adj001
densifloro.Adj001
densifoliado.Adj001
denso.Adj001S1
dentado.Adj001
dental.Adj111
dentário.Adj001
denteado.Adj001
denticórneo.Adj001
denticulado.Adj001
denticular.Adj105
dentiforme.Adj101
dentífrico.Adj001
dentígero.Adj001
dentilabial.Adj111
dentípode.Adj101
dentirrostro.Adj001
dentolabial.Adj111
dentolingual.Adj111
dentudo.Adj001
denunciado.Adj001
denunciador.Adj005
denunciante.Adj101
denunciativo.Adj001
denunciatório.Adj001
denunciável.Adj111
denutrido.Adj001
denutriente.Adj101
deodáctilo.Adj001
deontológico.Adj001
deoperculado.Adj001
deparador.Adj005
departamental.Adj111
depascente.Adj101
depauperado.Adj001S1
depauperador.Adj005
depauperante.Adj101
depenado.Adj001S6
depenável.Adj111
dependente.Adj101
dependurado.Adj001
dependurável.Adj111