Portuguese Verbal Conjugation: Factorizations and Algorithm

António Horta Branco
Tiago Castro Henriques
Faculty of Sciences, University of Lisbon

Introduction

Automatic verbal conjugation is a Natural Language Processing task by means of which the complete set of inflected forms of a verb together with associated inflectional features is provided. A restricted variant of this task consists in providing the inflected form of a verb corresponding to a certain set of feature values.

Automatic conjugation is a non-trivial task from a computational point of view inasmuch as it cannot be reduced to a simple lookup in a list of pairings of verbs and corresponding sets of inflected forms. As the lexicon slowly but steadily evolves with the addition of new verbs, an automatic conjugator has to be able to provide a correct output to previously untackled verbs, if complete and robust coverage of a given natural language is intended.

The feasibility of a conjugator relies on the fact that, in spite of the conjugational irregularities exhibited by the most common verbs, the conjugation of neologisms adheres to the overall network of regularities governing conjugation. Therefore, the major engineering constraints in the development of such a tool consist in both obtaining an exhaustive record of irregularities and in designing an accurate specification of the regularities that should be adhered to.

If efficiency is a concern, as it certainly has to be the case here, an additional requirement is that the system of regularities be kept as concise and practical as possible, even if this implies that the rules being considered do not take into account all the wealth of sophistication accumulated by analyses driven by morpho-phonological, psycholinguistic, etymological or other sorts of concerns. This does not mean that such analyses do not provide the fundamental guidelines for drawing the relevant specification. By the same token, this does not imply either that the exhaustiveness of the specification achieved may not be an important contribution to uncover less studied issues or inspire new analytical results.

In this paper, we report on an algorithm we designed for automatic verbal conjugation, with special focus on the system of regularities underlying its workings.

In Section 2, we describe the major regularities of verb conjugation, while in Section 3; we describe the various sub-regularities of so-called irregular verbs. The irreducible irregularities of irregular verbs are discussed in Section 4. Section 5 details how a lexicon of verbs can be organized that takes into account the

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística,* Lisboa, APL, 2002, pp. 201-209.**

regularities specified in the previous Sections and how it can be of practical use for our algorithm.

Finally, in Section 6, we briefly describe the algorithm.

Major regularities

The so-called regular verbs of Portuguese can be grouped into three conjugations according to which of three sets of terminations they use. The first conjugation contains verbs whose infinitive ends with *–ar*: This is currently the only productive conjugation since all new verbs introduced into the language fall into this conjugation.[1] As for the second and third conjugations, they contain verbs whose infinitive ends in *–er*, and *–ir*, respectively.[2]

Each conjugation contains 76 forms, corresponding to:

- 10 tenses with 6 forms each:

Presente Indicativo (PresInd), Presente Conjuntivo (PresConj), Pretérito Perfeito Indicativo (Perf), Pretérito Imperfeito Indicativo (ImpInd), Pretérito Imperfeito Conjuntivo (ImpConj), Futuro Indicativo (FutInd), Futuro Conjuntivo (FutConj), Pretérito mais-que-perfeito (MqPerf), Condicional (Cond) and Infinitivo Pessoal (InfPess)

- 2 tenses with 5 forms each:

Imperativo Afirmativo (ImperAfirm) and Imperativo Negativo (ImperNeg)

- 1 tense with 4 forms:

Particípio Passado (Part)

- 2 other forms:

Gerúndio (Ger) and Infinitivo (Inf)

The complete set S_V of inflected forms of a regular verb V is fully predictable.[3] To construct S_V, one just needs to: *i)* obtain the nucleus of V by removing the last

[1] It may be interesting to note that in a handbook for Portuguese verbal conjugation such as the one by Gramado, with a list of around 12,000 verbs, 80% belong to the first conjugation, 5% to the second conjugation, and 3% to the third conjugation. This leaves about 12% of the listed verbs in the broad category of irregular verbs (Gramado, N., 2001, *Dicionário de Verbos Portugueses*, Plátano Editora).

[2] We will be using the following abbreviations: *Conj-ar*, *Conj-er* and *Conj-ir* for the first, second and third conjugations, respectively.

[3] To clarify the terms that we will be using, the *nucleus* of a regular verb is defined as the longest initial string of letters that is shared by the forms of that verb, e.g. every form of *agradar* has *agrad-* as its nucleus. The *termination* of a form is the final string of characters that is left when the nucleus of that verb is removed. The *mediator* of a tense of a verb is the longest initial string shared by most of the terminations of that tense of the verb. The *terminator* of a form is the final string that is obtained by removing the mediator from the termination.

two letters of its infinitive form; *ii)* take the set S_P of inflected forms of a regular verb *P* of the same conjugation of *V*, conventionally taken as its standard representative; and *iii)* in S_P, systematically replace the nucleus of *P* with the nucleus of *V* to obtain S_V.

The tenses *Ger* and *Inf* are in a sense trivial, since there is only one possible termination for each regular conjugation. *Ger* has terminations *-ando*, *-endo* and -*-indo* for *Conj-ar*, *-er* and *-ir*, respectively, while *Inf* has terminations *-ar*, *-er* and -*-ir* for *Conj-ar*, *-er* and *–ir*.

The tenses *ImperAfirm* and *ImperNeg* are obtained in quite straightforward fashion as well, since all of their forms are identical with forms from two other tenses: *ImperNeg* from *PresConj*, and *ImpAfirm* from *PresInd* and *PresConj*.[4]

By deducting from the 76 forms the forms belonging to *ImperAfirm*, *ImperNeg*, *Ger* and *Inf*, there are yet 64 forms of a regular verb to be derived.

a. Terminations

If one considers that any regular verb of a given conjugation is as representative of that conjugation as any other, each conjugation can be simply represented as the set of pairings of inflected features and terminations.

For each of the tenses *PresConj*, *ImperNeg*, *FutInd*, *FutConj* and *InfPess* of any conjugation plus *Perf* of *Conj-ir*, the terminations of its forms have a common initial string, i.e. a mediator. For instance, every termination of *FutInd* of *Conj-ar* begins with *-ar-*.

For the tenses *MqPerf*, *Cond*, *ImpConj* and *ImpInd* of a given conjugation, the terminations of its forms also have a mediator – e.g. *-ess-* for *ImpConj* of *Conj-er* – and additionally two specific forms, viz. *1P* and *2P*,[5] require an accent on the last vowel of the mediator – e.g. *-êss-* for *1P* and *2P* of *ImpConj* of *Conj-er*.

Contrarily to the above tenses, in *PresInd* and *Perf*, a few terminations do not share a common initial string. In *PresInd* of *Conj-ar* and *Conj-er*, there is an initial string common to all terminations except for *1S*, where it is replaced by *–o*. In *PresInd* of *Conj-ir*, the forms *1P* and *2P* share a common initial string, namely *–i-*, while the terminations of the remaining forms result from some replacement of that string.

As for *Perf* of *Conj-er*, only the termination of *1S* does not share a common initial string with the other terminations of this tense, while in *Perf* of *Conj-ar*, this happens with the terminations of *1S* and *3S*. Additionally, the latter also requires an accent on the first letter of the mediator.

These factorizations are summed up in the following table, where one can count 12 different mediators:

[4] This is valid with the single exception of the verb *ser* in the forms *2S and 2P* and of the verb *haver* in the form *2P* of *ImperAfirm*.

[5] We will be using the following abbreviations: 1, 2 and 3 for first, second and third person, respectively; S and P for singular and plural, respectively.

Conj-*ar*	**Conj-*er***	**Conj-*ir***	**Tenses**
-e-	-a-	-a-	PresConj
-e-	-a-	-a-	ImperNeg
-ar-	-er-	-ir-	FutInd
-ar-	-er-	-ir-	FutConj
-ar-	-er-	-ir-	InfPess
-ar-	-er-	-ir-	Cond
-ar- 1P2P/-ár-	-er- 1P2P/-êr-	-ir- 1P2P/-ír-	MqPerf
-ass- 1P2p/-áss-	-ess- 1P2P/-êss-	-iss- 1P2P/-íss-	ImpConj
-av- 1P2P/-áv-	-i- 1P2P/-í-	-i- 1P2P/-í-	ImpInd
-a- 1S/-o	-e- 1S/-o	-e- 1S/-o, 1P/-imos, 2P/-is	PresInd
-a- 1P/-á- 1S/-ei, 3S/-ou	-e- 1S/-i	-i-	Perf
-ad-	-id-	-id-	Part

Table 1 – Mediators

For every form of any conjugation, almost all terminations are obtained by combining the 12 mediators above with the following terminators, grouped into 7 sets:

1S	**2S**	**3S**	**1P**	**2P**	**3P**	**Tenses**
-	-s	-	-mos	-is	-m	PresInd, PresConj and ImperNeg
-e	-es	-e	-emos	-eis	-em	ImpConj
-a	-as	-a	-amos	-ais	-am	ImpInd, MqPerf
-ia	-ias	-ia	-íamos	-íeis	-aim	Cond
-ei	-ás	-á	-emos	-eis	-ão	FutInd
-	-es	-	-mos	-des	-em	InfPess, FutConj
-i	-ste	-u	-mos	-stes	-ram	Perf

Table 2 – Terminators

The terminations that escape this factorization are the *1SPresInd* of the three conjugations (*-o*), *1SPerf* of *Conj-ar* and *Conj-er* (*-ei* and *-i*, resp.), *3SPerf* of *Conj--ar* (*-ou*), *1P* and *2P* of *Conj-ir* (*-imos* and *-is*, resp.), as can be seen from the last lines of the rows for *PresInd* and *Perf* in Table 1.

Taking into account these overriding cases and the accent rules, also displayed in the second line of the rows in Table 1, the combination of the 12 mediators (Table 1) and the 42 terminators (Table 2) yields the (3x64=) 192 inflected forms of the three conjugations.

Pseudo-irregularities

There are verbs that are regular from a phonological point of view, though they do exhibit written forms not complying with the regularities specified above.

This happens with verbs whose last string of the nucleus is:

i. a letter that cannot precede every vowel: That is the case of –ç-. For instance, the verb *dançar* has the *1SPresInd* form *danço* and the *1SPresConj* form *dance*.

ii. a single letter representing more than one phonological value: That is the case of *-c-* and *-g-*, which when followed by *a*, *o* or *u*, represent [k] and [g], respectively, while when followed by *e* or *i*, represent [c] and [3], respectively. For instance, the verbs *buscar* and *pagar* have the *1SPresInd* forms *busco*, and *pago*, and the *1SPresConj* forms *busque* and *pague*.

iii. a sequence of two letters representing more than one phonological value: That is the case of *-gu-* and *-qu-*, which when followed by *a* or *o*, represent [gw] and [kw], respectively, while when followed by *e* or *i*, represent [g] and [k], respectively.[6] For example, the verb *enxaguar* has the *3SPresInd* form *enxagua* and the *3SPresConj* form *enxagúe*; the verb *erguer* has the *1PresInd* form *erguemos* and the *1SPresInd ergo*; the verb *pagar* has the *1SPresInd* pago and *1SPerf paguei*. On the other hand, the verb *adequar* has the *1PPresInd adequamos* and the *1PPresConj adecue*[7]; the verb *extorquir* has the *1PPresInd* form *extorquimos* and the *1SpresInd* form *extorco*[7].

Pseudo-regularities are observed also in verbs whose nucleus terminates in two adjacent vowels (optionally followed by a consonant) that have to be identified not as diphthongs but as forming a hiatus. In such cases, the last vowel is marked with an accent. For example, the verbs *azougar* and *aceitar* have *1SPresInd* forms *azougo* and *aceito*, while the verbs *saudar* and *arruinar* have *1SPresInd* forms *saúdo* and *arruíno*.[8]

Sub-regularities

Many of the so-called irregular verbs, though they deviate from the global patterns exhibited by the verbs fully adhering to the three regular conjugations, are

[6] Exceptions to this are lexemes like the verbs *arguir* or *aguentar*, or the nouns *equidade* or *frequente*.

[7] Many conjugation handbooks prefer to assign a gap for these forms.

[8] Verbs ending in *-oiar*, e.g. *boiar*, *comboiar*, *engoiar*, etc., excluding *apoiar*, exhibit an analogous difference but due to a change in nature of the vowel. For instance, for the verb *boiar*, *1PPresInd* form is *boiamos* while *1SPresInd* is *bóio*.

not completely opaque with regard to possible factorizations. Non-regular verbs may exhibit what can be termed as sub-regularities, in the sense that they adhere to a larger network of less comprehensive regularities.

For non-regular verbs, sub-regularities may be found on the nucleus, on the terminations or on both.

c. Nuclei

Taking the nuclear part of the verbal forms, we focus on the tense nucleus, i.e. the longest initial string of characters that is shared by the inflected forms of the same simple tense of a verb. Regular verbs can be characterized as having the same tense nucleus for every tense, in which case all tense nuclei are identical to the verb nucleus. For instance, the verb *agradar* has *agrad-* as its indicative present tense nucleus; as this is a regular verb, *agrad-* is the tense nucleus in all tenses.

When one considers the tense nuclei across a comprehensive set of irregular verbs, a few patterns emerge.

First, a large number of irregular verbs do not have a single verb nucleus. As an example, for the verb *dizer*, the nucleus of the present indicative is *diz-*, while the nucleus of the future indicative is *dir-*.

Second, for any poly-nuclear verb, there are at most four different tense nuclei. For example, the verb *equivaler* has two tense nuclei, viz. *equival-* and *equivalh-*; the verb *poder* has three tense nuclei, viz. *pod-*, *poss-* and *pud-*; and the verb *fazer* has four, viz. *faz-*, *faç-*, *fiz-* and *far-*.

Third, there are at most five partitions of tense nuclei across poly-nuclear verbs. Within each of the following group of tenses, every tense has the same tense nucleus if there are no singularities or gaps (cf. Section 4):

0 *Ger*, *Part*, *InfPess* and *PresInd* (except *1S*) and possibly any other tense
A *PresConj* and *1SPresInd*
B *Perf*, *MqPerf*, *ImpConj* and *FutConj*
C *ImpInd*
D *FutInd* and *Cond*

Fourth, not all the above partitions may co-occur. One observes at most five possible combinations, illustrate here with five examples:

0A	*medir*:	0/*med-*, A/*meç-*
0AB	*saber*:	0/*sab-*, A/*saib-*, B/*soub-*
0ABC	*vir*:0/*v-*, A/*venh-*, B/*vi-*, C/*vinh-*	
0ABD	*dizer*:	0/*diz-*, A/*dig-*. B/*diss-*, D/*dir-*
0B	*dar*:	0/*d-*, B/*de-*

Fifth, the tense nucleus of the tenses in partition 0 is the default tense nucleus in the sense that is holds for every tense whenever any of the other four tense nuclei A, B, C and D do not hold.

d. Terminations

Terminations may also deviate from the global patterns observed for the terminations of regular verbs in accordance with some sub-regularities, though to a more limited extent than what can be detected for tense nuclei.

The verbs for which some terminations deviate from regular patterns according to the same sub-regularity bear the same termination for the infinitive form.

For verbs ending in *-ear*,[9] the regular terminations of

- *1S*, *2S*, *3S* and *3P* of *PresInd* are overridden by *-io*, *-ias*, *-ia* and *-iam*, respectively. For example, these *PresInd* forms of verb *passear* are *passeio/*passeo*, *passeias/*passeas*, *passeia/*passea*, *passeiam/*passeam*
- *1S*, *2S*, *3S* and *3P* of *PresConj* are overridden by *-ie*, *-ies*, *-ie* and *-iem*, respectively. For example, these *PresInd* forms of verb *semear* are *semeie/*semee*, *semeies/*semees*, *semeie/*semee*, *semeiem/*semeem*

For verbs ending in *-zer* or *-uzir*, the regular termination of *3S* of *PresInd* is deleted. For example, this form of verb *jazer* is *jaz/*jaze*, and of verb *produzir* is *produz/*produze*.

Non-regularities

Some individual forms escape any common pattern, be it a major regularity or a sub-regularity. Such non-regularity can be found in the fact that there are forms which are not used or which are singular.

e. Gaps

Verbs for which not all the 76 expected forms are used are usually termed as defectives.

Some forms may be not used because they refer to non-existing state of affairs in the current world. That is the case of *1S*, *2S*, *1P*, *2P* and *3P* forms of meteorological-like verbs, such as *chover*, *nevar*, etc., and the case of *1S*, *2S*, *1P* and *2P* forms of verbs for animal sounds or actions, such as *balir*, *galopar*, *marrar*, etc.

Some other forms may not be in use because they do not sound euphonic – such as the form *1SPresInd falo* of verb *falir*, etc. – or due to some other undetermined reason.

There are five patterns of defectiveness:

D1 *1SPresInd*, *PresConj*
D2 *1S*, *2S*, *3S*, *3P PresInd* and *PresConj*
D3 all forms except *Ger*, *Part*, and *3S*, *3P PresInd* and *ImpInd*
D4 all forms except *Ger*, *Part*, and *3S* of every tense
D5 all forms except *Ger*, *Part*, and *3S* and *3P* of every tense

[9] Verbs ending in *-iar* are arranged into three groups: some of them follow the same sub-regularity of verbs ending in *-ear*; some others are regular verbs; and still others admit both patterns.

f. Singularities

Some irregular forms owe their singularity to the elision of the nucleus, such as the forms *2SPresInd vais* of the verb *ir*, *3SPresInd é* of verb *ser*, etc.

Some other irregular forms just result from an idiosyncratic termination, such as the forms *1SPerf vim* of verb *vir*, *1SPreInd sei* of verb *saber*, etc.

Many irregularities are due to irregular *Part* forms, such as the form *aberto* of the verb *abrir*, or to alternative *Part* forms, such as the form *liberto* (reg.: *libertado*) of the verb *libertar*.

Data repositories

The information on the conjugational characteristics of Portuguese verbs can be recorded in three major data repositories.

On the one hand, the irregular forms are collected in a list of singularities. Each entry in this list includes the irregular form, its inflectional features and the corresponding verb.

On the other hand, the information on the regular forms of a verb is registered in the lexicon of verbs. Each entry of this lexicon corresponds to a verb and includes

- its infinitive form
- the code of the type of defectiveness (D1-D5), if any
- the accentuation specificities for adjacent vowels (cf. Section 2.2), if any
- its model

A model is a prototypical verb that represents a set of verbs that share the same regularities, sub-regularities and singularities. Given the factorization described above, we isolated 49 models whose details are recorded in a third repository, the list of models.

Algorithm

To generate the set of forms of a verb from its infinitive form V, the following steps are performed:

```
i.  pre-process V so that all context-restricted consonants or digraphs are
    rewritten using special purpose, context-independent symbols[10]
ii. look up V in the lexicon of verbs
    o if V is found, take its model verb M, and possible accentuation
      particularities and gaps
    o else determine the model M by examining the termination of the
      infinitive
```

[10] This makes it possible to process all verbs with this kind of consonants at the end of the nucleus as regular verbs.

```
iii. take the terminations for M
iv.  take the set of transformations (termination replacement and accentuation
     rules) for M, if any
v.   for each tense T:
     o find the tense nucleus MTN of M
     o find the tense nucleus VTN of V
     o for each inflection feature IF of T, look up M+IF in the list of irregular
       forms:
       ▪ if an entry E is found, we determine the termination T1 by removing
         MTN from E
         • add the termination T1 to the nucleus VTN to obtain form A1
         • associate form A1 with IF
       ▪ else:
         • look up the termination corresponding to IF in the set of
           terminations for M, and add it to VTN to obtain form A2
         • if any transformations apply for this particular IF, use apply them
           to A2 to obtain form B
         • else let form B = A2
         • associate form B with IF
vi.  perform the inverse transformation as in step i. to every form of V.
```