The processing of verbal inflection ambiguity: characterization of the problem space

António Branco, Francisco Costa and Filipe Nunes
University of Lisbon, Dept of Informatics
NLX- Natural Language and Speech Group

1. Introduction

Like for other Romance Languages, but in contrast with many other languages of the world, one of the most salient features of the Portuguese language concerns the complexity of its verbal inflection system. A well known trait of this complexity resides on the fact that, for a non defective verbal lemma, the corresponding conjugation table typically contains over 160 conjugated forms.

From the viewpoint of the processing of Portuguese, either by humans or by artificial agents, an important dimension of the difficulty posed by this system of verbal inflection arises not only from the large number of morphological affixes and rules involved, but also and above all from the processing costs incurred in the resolution of the ambiguity induced by verbal inflection.

Nevertheless, while the system of verbal rules and affixes has been largely studied and characterized (Cunha e Cintra, 1984, Mateus et al., 1990, Villalva, 2003 among others), very little if anything is known about the size and the characteristics of the problem space that the verbal processors have to deal with when tackling the inflection-driven verbal ambiguity. In this connection, many basic questions are waiting to be answered. For instance, questions such as: What is the inflection-ambiguity rate in the lexicon of verb forms? And in the verb forms occurring in actual text? From the verb forms that are inflection-ambiguous, what proportion corresponds to those that are ambiguous with respect only to the lemma? And how about those that are ambiguous with respect to the feature bundle expressed? What are the most used feature bundles? As for the verb forms that are ambiguous with respect to the feature bundles expressed, what is the relative frequency of each one of the possible readings? etc.

Answers to questions like these are of paramount importance in different regards. From a theoretical perspective, for instance, they will permit to know better the actual conditions of the usage of the Portuguese language and will contribute to foster progress in the characterization of the nature and complexity of the mental processes possibly involved in the processing of Portuguese. From a practical perspective, in turn, they will provide important insights for the design of more efficient pedagogical materials and devices either to teach writing skills for native speakers or to support the learning of

XXII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2007, pp. 157-168

Portuguese as a second language.[1] They will provide also valuable indicators for the characterization of the computational difficulty that verbal lemmatizers have to address and will help to devise better suited heuristics for the resolution of the sort of ambiguity at stake here.

This situation of lack of knowledge in this respect is to a large extent a consequence of the fact that, in order to answer questions like those above, one needs to resort to a set of linguistic resources and research tools that were not available so far. Such resources include machine readable lexicons and corpora accurately annotated with respect to inflection features, while the set of relevant tools includes automatic verbal conjugators, featurizers and lemmatizers. For the results reported in the present paper, we resorted to the resources and tools developed in the scope of the research initiatives undertaken in our group. They include a 1/4M token corpus, and an automatic featurizer and lemmatizer (Barreto *et al.*, 2006). They include also a conjugator (for online services and further documentation, see http://lxconjugator.di.fc.ul.pt and http://lxlemmatizer.di.fc.ul.pt).

In the present paper, our goal is thus to provide a first characterization of the problem space concerning the resolution of inflection-driven verbal ambiguity. In the next Section 2, the possible types of inflection-driven readings will be presented in detail. A characterization of the problem space, both in terms of the lexicon of verbal forms and in terms of the actual usage of verb forms in text, is provided in Section 3. Finally, in Section 4, concluding remarks are presented.

2. Preliminary issues

The processing of verbal forms in languages like Portuguese involves two procedures that take a verb form as input but have different post-conditions.

In one of the procedures, the intended output is the lemma corresponding to the input form, which is a designated verb form picked from the conjugation table to which the input form belongs. In Portuguese, the form assumed as the lemma is the Infinitivo Impessoal verb form, as exemplified below:

```
consumiu → consumiu / consumir                (1)
```

In the other procedure, the intended output is the bundle of values of the grammatical features encoded by the inflectional suffixes, as in the following example:

```
deu → deu # Indicativo, Pretérito Perfeito, 3rd, singular (2)
```

A verb form in Portuguese can bear suffixes that encode the values of up to 4 inflectional features from the following list of 6 admissible features and their values:

[1] For example, by selecting the top most frequent inflected forms, feature bundles or lemmas to occur in the textbooks, exercises or other pedagogical auxiliaries will maximize the number of linguistic situations where the learned skills can be put into use with respect to the number of learning hours needed to acquire them.

Mood (with 6 values): Indicativo, Conjuntivo, Imperativo, Infinitivo, Gerúndio, and Particípio Passado

Polarity (2): Afirmativo and Negativo

Tense (15): Presente, Pretérito Perfeito Composto, Pretérito Perfeito Simples, Pretérito Imperfeito, Pretérito mais-que-Perfeito Composto, Pretérito mais-que-Perfeito Simples, Pretérito mais-que-Perfeito Anterior, Futuro do Presente Simples, Futuro do Presente Composto, Futuro do Pretérito Simples (aka Condicional), Futuro do Pretérito Composto, Infinitivo Impessoal Presente, Infinitivo Impessoal Pretérito, Infinitivo Pessoal Presente, and Infinitivo Pessoal Pretérito.

Person (4): First, Second, Second courtesy, and Third.

Number (2): Singular, and Plural.

Gender (2): Masculine, and Feminine.

This second procedure has typically been termed in the Natural Language Processing (NLP) literature as morphological analysis, though no word analysis in terms of its constituency is intended, but only the extraction of feature values. A more accurate designation for this procedure is thus featurization, which we will adopt here.

The first procedure described above, in turn, has been termed as lemmatization.

Differently from what the two easy examples above may suggest, when performed upon any verb form occurring in context, these two procedures are computationally much less trivial than a mere look up in some correspondence table relating forms and bundles of feature values. As many verb forms yield more than one output, lemmatization and featurization in context are thus non trivial ambiguity resolution procedures. As the example below illustrates, lemmatization of a lemma-ambiguous verb form in the context of its occurrence has to decide which one of the possible lemmas is the appropriate one:

```
consumo → consumo / consumir                    (3)
          consumo / consumar
```

A similar consideration applies to the process of featurizing feature-ambiguous forms in their context of occurrence, as will be the case with the following form:

```
deram → deram # Indic, Pretérito Perfeito, 3rd pers, plural (4)
        deram # Indic, Pret mais-que-Perfeito, 3rd pers, plu
```

Interestingly, for some verb forms, the two procedures may colapse into a single one, as a correct decision on which lemma to pick in the context at stake implies a correct decision on which feature values to pick in that same context, or vice-versa. This circumstance occurs when a form may express two different sets of feature values, each one of them in correspondence with a different lemma, as exemplified below:

```
virei → virei / vir # Indic, Futuro, 1st, sing      (5)
        virei / virar # Indic, Pretérito Perfeito, 1st, sing
```

This is likely the reason behind the fact that the featurization procedure is sometimes loosely referred to in the literature under the general term of lemmatization, though no search for a lemma may eventually be involved. In this paper, we will adhere also to this practice of using "lemmatization" *lato sensu* when no confusion may arise.

The circumstance that two decisions – lemmatization *stricto sensu* and featurization – are required for verb forms, and that in some cases they may even be conflated into a single decision, brings eloquently to light that these forms are expressions with special semantic behavior, in as much as they are, so to speak, doubly semantically loaded.

On the one hand, while denoting a state of affaires, a verb may be ambiguous between the conveying of different relations between entities of the world, as in the following example:

```
bate → bate / bater # IndPres3sg {bater, vencer,..}   (6)
       bate / bater # IndPres3sg {bater, remexer,..}
```

Resolving this sort of ambiguity is a clear instance of the more general NLP task known as Word Sense Disambiguation (WSD), applied here to a specific Part-of-Speech (POS) class, the class of verbs. Under this perspective, and in as much as a verb form may express more than one sense because it underlies conjugated forms of different lemmas, lemmatization turns out to be part of the WSD task for verbs:

```
consumo → consumo / consumir # IndPres1sg {consumir,..} (3')
          consumo / consumar # IndPres1sg {consumar,..}
```

On the other hand, the ambiguity of a verb may be rooted not (only) in its lemma but in its bundle of suffixes. For instance, while denoting a given state of affaires, a verb may be ambiguous among the conveying of different temporo-aspectual relations between relevant events holding at the so called utterance time, reference time or event time, as the example above exemplifies:

```
deram → deram / dar # IndPretPerf3p                    (7)
        deram / dar # IndPretmqPer3p
```

This sort of ambiguity may emerge also in terms of Mood values,

```
dê → dê / dar # ConjuntivoPres3s                       (8)
     dê / dar # Imperativo2sCourtesy
```

in terms of Polarity values,

```
dêmos → dêmos / dar # ImpAfirmativo1p                  (9)
        dêmos / dar # ImpNegativo1p
```

in terms of Person values,

```
dava → dava / dar # IndPretImp1s                    (10)
       dava / dar # IndPretImp3s
```

in terms of Number value,

```
parti → parti / partir # IndPretPerf1Singular       (11)
        parti / partir # Imper2Plural
```

in terms of Gender,

```
assente → assente / assentar # ParticípioPassado3sMasculine (12)
          assente / assentar # ParticípioPassado3sFeminine
```

or in terms of several of these dimensions concomitantly.

In this sort of verb ambiguity, rooted in the expression of meaning by the inflectional terminations, the task of featurization consists in assigning to a verb form in context the appropriate tag, made of the combination of sub tags with values for Mood, Tense, etc. Under this perspective, and in as much as a verb form may express more than one feature bundle, lemmatization lends itself also to be envisaged as an extension of the more general POS tagging task for verbs.

Nevertheless, as discussed above in connection with (5), the ambiguity induced by the inflectional system may extend also to the realm of the basic relation expressed by the verb forms, and one can find forms with the two dimensions of ambiguity, lemma- and termination-driven. In these cases, the procedure of featurization appears also as part of the more general WSD task.

From this preliminary study, a few insights concerning the procedure of lemmatization emerge:

- It is an ambiguity resolution task;
- It has to handle expressions that may conflate several dimensions of ambiguity;

3. Problem space

In this section, we look for a better understanding of the impact of lemma- and inflection-driven ambiguity on verbal lemmatization.

3.1 The lexicon of forms with ambiguity for lemmatization

In order to undertake the study of the verbal lexicon, we relied on a verbal conjugator and a verbal lemmatizer. They are both fully fledged tools with exhaustive coverage, taking as input any verb form with orthographically well-formed termination, including possible neologisms. They are available online and documented in http://lxconjugator.di.fc.ul.pt and http://lxlemmatizer.di.fc.ul.pt.

We used a lexicon of lemmas consisting of 11 350 entries. These entries contain information about the inflectional behavior of the lemma, including details about defectiveness, double forms of past participles, blocked imperatives, alternative orthographic variants (European and American), etc.[2]

With the help of the conjugator, the conjugation tables for every lemma in this lexicon were generated. Typically, for a non defective verb, there are 168 different feature bundles (combinations of inflectional feature values) and consequently the same amount of conjugated forms in its fully-fledged conjugation table, with this figure raising to 172 in the case of unaccusative, transitive and ditransitive verbs, with Past Participle forms inflecting for Gender and Number.

A lexicon of conjugated forms results from collecting all the non compound forms[3] in the conjugation tables for every item in the lexicon of lemmas. This lexicon contains every conjugated form of any of the original list of lemmas, where such a conjugated form consists of a verb form associated with the corresponding feature bundle and lemma. This lexicon consists of 816 830 such entries.[4]

The analysis of this lexicon of conjugated forms permits some interesting advances in the understanding of the inflection-driven verbal ambiguity.

3.1.1 Ambiguity inside conjugation tables

There are ambiguities that affect almost every conjugated table in a systematic fashion. With the exception of only a few highly defective verbs, each lemma yields several ambiguous verb forms, that is they underlie several conjugated forms, bearing different feature bundles. To refer just an example, in every conjugation table, the conjugated forms expressing the Indicative, Pretérito Perfeito, 3rd Person, Plural and the Indicative, Pretérito mais-que-Perfeito, 3rd Person, Plural have identical verb forms (e.g. amaram, beberam, partiram). Interestingly, in a typical conjugation table, with 168 conjugated forms, 82 of them are underlaid by 23 ambiguous verb forms. This implies that only around 1/2 of the items in a typical conjugated table are not affected by this systematic ambiguity among forms of the same lemma.

3.1.2 Ambiguity across conjugation tables

Verb forms are not ambiguous only between conjugated forms of the same lemma, i.e. of the same conjugation table. As discussed above and illustrated in (3'), verb forms may be ambiguous because they pertain to conjugated forms that have both the same verb form and the same associated feature bundle, but belong to different conjugation

[2] For the sake of simplicity, 284 additional lemmas corresponding to verbal entries with inherent clitic were ignored.

[3] Note that for the sake of lemmatization, processing compound forms resorts to processing the simple forms of the auxiliary verbs.

[4] Conjugated forms of 2nd person of courtesy, with systematic ambiguity with the 3rd person of the same Tense, were not added to this figure.

tables, i.e. they are inflected forms of different lemmas. The analysis of the lexicon of conjugated forms permits to know that it contains 280 items in such circumstances, which correspond to 141 verb forms displaying lemma-only ambiguity.

Besides, other verb forms may be ambiguous on both counts, i.e. they are both lemma- and termination-ambiguous. An extreme example of this situation is found in (5), where a verb form is just two-fold ambiguous and each one of the two conjugated forms has both a different lemma and different feature bundle associated to it.

While there are 184 verb forms under this strict two-fold, lemma- and termination--ambiguity, the majority of ambiguous verb forms that pertain to conjugated forms belonging to different conjugation tables belong also to conjugated forms of the same conjugation table, as the following example illustrates:

```
doa → doa / doar # IndPres3s          (13)
      doa / doar # ImpAfirm2s
      doa / doer # ConjPres3s
```

There are as much as 886 verb forms under this less strict circumstance, which added to the previous figure reveals that our lexicon of conjugated forms contains 1 070 verb forms that are ambiguous both inside and across conjugation tables, that is they are lemma- and termination-ambiguous.

Collecting the above figures into the table below and completing it, it is possible to get a quantitative synopsis of the different sorts of lexical ambiguity lemmatization has to cope with.

Ambiguity root	Examples	# Verb forms
lemma-only	(3)	141
lemma-and-termination	(5), (13)	1 070
termination-only	(7)-(12)	159 376
total		160 587

Table 1. Verb form tokens in the lexicon by type of lemmatization ambiguity

3.1.3 Lexical ambiguity ratio

Getting back to the lexicon of conjugated forms, we can now obtain the lexicon of verb forms, to which each verb form, ambiguous or not, contributes only one entry.

It is possible to divide the lexicon of conjugated forms into conjugated forms with and without underlying ambiguous verb forms. As explained above, this lexicon contains a total of 816 830 items, of which 438 064 count as having no underlying verb forms that are ambiguous from the lemmatization point of view. This implies that these conjugated forms contribute with an identical amount of entries to the lexicon of verb forms.

Together with the total from Table 1, we thus know that the lexicon of verb forms has a total of 598 651 items, of which over 1/4 are ambiguous from the perspective of

lemmatization *latu senso*, and about the same figure from the perspective of featurization as just 114 forms are lemma-only ambiguous.

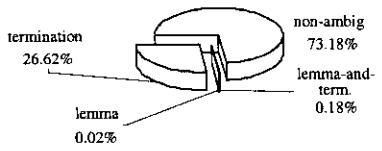

Fig. 1. Lexicon of verb forms

Taking into account that the 598 651 verb forms underlie 816 830 conjugated forms, the resulting lexical ambiguity ratio is 1.36 from the viewpoint of lemmatization *lato sensu*, and about the same ratio from the viewpoint of featurization, with 598 931 pairs of verb forms and lemmas underlying the lexicon of conjugated forms.

Further insight into the structure of the lexicon of verb forms from the perspective of ambiguity for lemmatization is obtained in the following chart, where the lexicon was partitioned into subsets containing forms with identical degree of ambiguity, whose size is displayed in the Y axis:

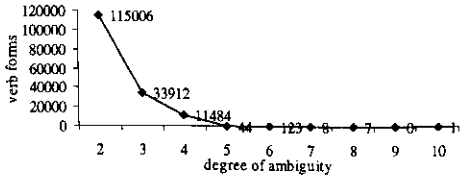

Fig. 2. Verb forms in the lexicon per degree of ambiguity

3.2 Usage of forms with ambiguity for lemmatization

Turning now to the characterization of the usage of the lexical resources identified above, a first step consists in determining how much of these resources are typically put to use and what is the impact of ambiguity for lemmatization of actual texts. In order to accomplish this, we used a corpus with 261 385 tokens accurately hand annotated with respect to POS and verbal inflection values, of which 13.51% (35 306) are verbal tokens.[5] This corpus includes texts from newspapers (ca. 3/5) and fiction (Barreto *et al.*, 2006).

[5] Compound tenses (with 63 feature bundle types) were counted as contributing two tokens.

3.2.1 Lexical resources used

Grouping the verbal tokens in the corpus by types, and sorting them by analytical categories, several figures indicative of the usage of verbal lexical resources obtain:

Verbal types	Lemmas	Feat. bundles	Conj. forms	Ambig. forms
# in the lexicon	11 350	109[4]	816 830	160 587
# in the corpus	1 951	82	8 635	4 142
Usage rate	17.19%	75.23%	1.06%	2.58%

Table 2. Usage rate of verbal lexical resources in a 1/4M corpus

The verbal lemmas put to use are just ca. 1/6 of the lemmas available in the lexicon, which reflects the limited size and above all the scarce genre diversity of the working corpus. This did not hampered however that over 3/4 of the relevant combinations of verbal inflection feature values were put to use. Interestingly, only a tiny portion of the lexically possible conjugated forms, involved or not in ambiguity, occur.

3.2.2 Ambiguity ratio

To get a deeper insight into the ambiguity for lemmatization, significant indicators were collected below, with the distribution of tokens with different sorts of ambiguity:

Ambiguity root	Examples	# Tokens
lemma-only	(3)	695
lemma-and-termination	(5), (12)	1 807
termination-only	(7)-(11)	15 063
total		17 565

Table 3. Verb forms by type of ambiguity for lemmatization

This helps us know that, from the 35 306 verbal tokens occurring in the corpus, 1/2 display ambiguity that calls to be solved by the lemmatization procedure:

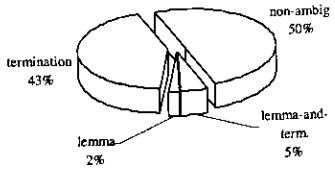

Fig. 3. Ambiguity of verb forms in corpus

Although only little more than 1/4 of the lexicon of verb forms contains ambiguous items (Fig. 1), this indicates that as a group they have an usage rate above average.

Note that there are 21 883 verb form types in the corpus. Accordingly, since they underlie the 35 306 conjugated verb tokens also in the corpus, the ambiguity ratio exhibited by the corpus turns out to be 1.61.

3.2.3 Further insights into verbal ambiguity

Taking into account the distribution of the frequency of verbal tokens, one finds that the 10 more frequent conjugated forms are 17.30% of the verbal tokens in the corpus (Tab. 4), while the 10% more frequent ones cover 84.95%. Ranking the verbal tokens by decreasing frequency, a Zipfian profile emerges (Fig. 4, right).

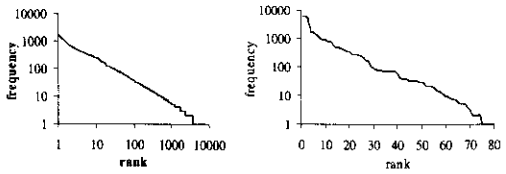

Fig. 4. Freq. of decreasingly ranked verb tokens (left) and feature bundle tokens (right)

As for feature bundles, 82 are instantiated in the corpus,[6] and the 10% more frequent ones occurs in 69.60% of the verb tokens in the corpus. The chart in Fig. 4 is obtained when feature bundles are ranked according to decreasing frequency in the corpus.

Conjugated form	#
é # IndPres3s	1762
foi # IndPretPerf3s	709
ser # InfImp	497
há # IndPres3s	409
está # PresInd3s	389
era # IndPretImp3s	329
são # IndPres3p	310
tem # IndPres3s	282
vai # IndPres3s	255
disse # IndPretPerf3s	248

Feature bundle	#
IndPres3s	6174
InfImpessoal	6007
IndPretPerf3s	5216
PartPassMs	1732
IndPres3p	1702
IndPretImp3s	1431
ParPassFs	1192
InfPess3s	1041
IndPretPerf3p	949
IndPres1s	920

Verb form	#
é	1762
foi	709
ser	564
há	409
está	391
ter	362
tem	354
era	341
são	310
disse	281

Table 4. The 10 more frequent conjugated forms, feature bundles and verb forms in the corpus

Also interesting for the purposes of lemmatization is to further detail the analysis of ambiguity for each verb form. In the corpus, a verb form type happens to underlie at most 4 different conjugated form types. Conjugated forms can then be ranked 1-4, according to their decreasing relative frequency in the corpus, with respect to the group

[6] Keeping in line with the analysis of the lexicon of conjugated forms, the tokens of 2nd person courtesy, with systematic ambiguity wrt 3rd person, contributed for the counting of the latter.

of conjugated forms underlaid by the same verb form type. Interestingly, it emerges that the tokens of the most frequent conjugated form of every verb form occurring in the corpus amount to 93.98% of the total verb tokens in the corpus:

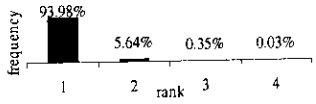

Fig. 5. Distribution in corpus of frequency ranked conjugated forms wrt verb form

4. Discussion

To the best of our knowledge, the present study on the quantitative linguistics of the verbal inflection system of Portuguese is the first of its kind.

Given the results obtained in the study of the problem space undertaken in the previous Sections, in particular those displayed in Fig. 5, the heuristic of the most frequent sense offers itself also as very promising WSD baseline method for verbal featurization. Accordingly, we performed an experiment with this heuristic. It turns out that a verbal lemmatizer that assigns to a given verb form in a text its most frequent feature bundle, when evaluated on 10% of the corpus held out from the learning phase, shows an F-measure[7] of 85.19%.

This result reveals that this quite simple heuristic displays a performance that can be deemed as very good when considered in the setting of present state-of-the-art WSD performance (Pedersen & Mihalcea, 2005).

Besides, this preliminary result suggests also that an algorithm based on a first option for the most frequent possible reading, with possible backtraking in case of incompatibility with discourse context, is an hypothesis for an efficient mental processing worth considering in further research addressing the processing of verbal inflection with other research methods and experimental tools.

References

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes & João Silva (2006) "Open Resources and Tools for the Shallow Processing of Portuguese", *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006), Paris: ELRA.

Cunha, Celso & Lindley Cintra (1984) *Nova Gramática do Português Contemporâneo*, Lisboa: Sá da Costa.

[7] F-measure= (2 x Precision x Recall) / (Precison + Recall), where Recall = R / T and Precision = R / A, with R the number of correctly resolved tokens, T the total number of tokens to be resolved, and A the number of tokens whose resolution was attempted .

Mateus, M. H. Mira, Amália Andrade, Maria do Céu Viana & Alina Villalva (1990) *Fonética, Fonologia e Morofologia do Português*, Lisboa: Universidade Aberta.

Pedersen, Ted & Rada Mihalcea (2005) *Advances in WSD*, Tutorial at ACL 2005, Association for Computational Linguistics.

Villalva, Alina (2003) "Estrutura Morfológica Básica", In Mateus *et al.*, *Gramática da Língua Portuguesa*, Cap. 22., Lisboa: Caminho.