On the problems of creating a golden standard of inflected forms in Portuguese

Diana Santos, Anabela Barreiro

Linguateca at SINTEF ICT
Pb 124 Blindern, 0314 Oslo, Norway
Diana.Santos@sintef.no, Anabela.Barreiro@label.ist.utl.pt

Abstract

This paper describes our attempt to build a consensus round the morphological analysis of a set of forms for Portuguese, to be used as a basis for the creation of a "golden list" in the first Morpholympics for Portuguese, *Morfolimpíadas*, an evaluation contest on Portuguese morphological analysis. This golden standard was used to rank participating morphological systems according to precision and coverage. The discussion in the paper is centered on the general choices made and the problems encountered. The paper ends with a short description of the (publicly available) resource.

Introduction

The goal of this paper is to present, to a larger community, the issues involved in creating an evaluation resource for morphological analysers (*lista dourada*, henceforth golden list), a set of correct analyses of Portuguese inflected forms, for the first *Morfolimpíadas* for Portuguese.

Evaluation contests for Portuguese

One of the aims of Linguateca, a distributed resource center for Portuguese NLP, is to organize evaluation contests so that it is possible to assess the state of the art, measure progress, and foster collaboration, or at least awareness, between several groups working in the same domain. So we organized AvalON (Santos & Rocha, 2003) as a first step to develop a tradition of evaluation in the language engineering community dealing with Portuguese.

Morfolimpíadas was the first evaluation contest in the scope of AvalON and had as its goal to compare current morphological analysers for Portuguese. Following one of the guiding principles of Linguateca, a common contest for all variants of Portuguese was organized (care was taken to have an equal representation of the major variants – from Brazilian and Portugal).

Morfolimpíadas

As it is customary with evaluation contests (see Hirschmann (1998) for a description of the evaluation model), we first ran a trial (described in Santos, Costa & Rocha, 2003), where a preliminary golden list was produced, for purposes of clarification and exchange of ideas among participants.

One of the main inspirations was Hausser's (1996) Morpholympics for German, although we ended up following his suggestions for further competitions, rather than the actual setup. Also, to avoid the complex issue of finding a board of authoritative judges, no qualitative evaluation was included.

The contest took place remotely over the Web in April 2003, with the final results being presented at a special session of Avalon'2003, on June, the 28th 2003. Seven systems participated (in alphabetical order: br.ispell, Jspell, LabEL-Intex, PALMORF, ReGra, Rsufixos, and Smorph-Pasmo), three from Portugal, three from Brazil and one, the winner (PALMORF), from Denmark.

Although the main emphasis was on morphology, we also ran comparisons of the systems as spell checkers and stemmers, when appropriate. In the present paper we are only concerned with the creation and use of the resource used as a golden standard. Other issues, such as the choice of the texts, the tokenization differences, and the different aims of the systems, are discussed elsewhere.

The concrete golden list

We present some examples from the golden list now, so that the reader can appreciate the nature of the resource involved in the discussions below. The complete resource, as well as the (anonymized) outputs of all participating systems, is publicly available on the Web.[1]

```
celular÷ADJ÷celular÷.÷S÷.÷I÷.÷.÷.
celular÷SUB÷celular÷.÷S÷.÷M÷.÷.÷bras

frete÷SUB÷frete÷.÷S÷.÷M÷.÷.÷.
frete÷V÷fretar÷PR_C÷S÷1÷.÷.÷.÷.
frete÷V÷fretar÷PR_C÷S÷3÷.÷.÷.÷.

Chaves÷PROP÷Chaves÷.÷S÷.÷.÷.÷.÷.
Chaves÷SUB÷chave÷.÷P÷.÷F÷.÷.÷.
```

Figure 1: Forms in the golden list: *celular*, *frete*, *Chaves*

From Figure 1 it can be seen that, for every form, a list of classifications was provided, in a shorthand form (not identical with the conventions of any of the systems), having values for form, PoS, lemma, tense, number, person, gender, degree, diminutive/augmentative, and "others". The value "." indicates that the feature is not relevant (like tense for nouns); the value "I" (indeterminate) indicates that although relevant, it is not defined (like gender for some adjectives).

In the last field (the only one with different semantics from the output of the systems), the golden list compilers were supposed to inscribe meta-information that might be relevant for finer evaluation: variant, relative frequency compared to the other analyses, stylistic information such as common error, foreign word, made-up form, etc.

General principles

[1] See http://www.linguateca.pt/Morfolimpiadas/.

The current paper describes several problems associated with the need to have a consensual reference resource that can be employed to compare different systems, created independently by different groups and under different theoretical assumptions. Such a need brings several kinds of constraints:

1. Only common dimensions can be compared
2. The resource must be fair relative to all systems
3. The metrics have to be made flexible in order to study the result of taking into consideration different details
4. The resource should reflect interesting cases

In addition, there must be some consensus among the annotators, and ideally the consensus should cover a larger set of forms than the ones included in the golden list.

Of course, every concern mentioned is debatable:

1. One might have aimed at a repository of right and desirable features, even though only one system (or even none) featured it. The problem here was to achieve an ideal object, instead of a concrete, goal-oriented tool for the current (and first) competition, in a convenient time frame, through e-mail communication.
2. One might as well give more weight to some features than to others and, therefore, at once put the systems in different classes (the very good, the middle and the basic ones). Again, for the need of cooperation and friendly treatment of participants, we decided not to do that, leaving such decisions – if practicable – for future competitions.
3. Although probably the most consensual consideration, the formulation above entails considerable vagueness. Which, of all possible parameters, should be tried and weighted, and how to implement a fair smoothing of the factors involved? Probably since we showed a considerable care in treating the idiosyncrasies of all systems involved – so that the "spirit of the system" was preserved –, the participants trusted the organization and let us try out several different computation functions and decide on the final ones.
4. The last desideratum is – astonishing as it may be for an unprepared reader – the least defensible. In fact, in real systems, "interesting" problems tend to correspond to rare cases and be therefore statistically irrelevant. So, weighting a golden list by frequency2 – with considerable addition of "uninteresting" cases – could have been preferable.

This last option was, in a way, conditioned by the creation process. In fact, an initial golden list was created cooperatively by the trial participants according to their interests; therefore, it was not feasible to ask for frequency-weighted sets of forms. In addition, to increase discrimination, to the final golden list we added some forms among which there was actual disagreement between the participating systems.

The creation process

Considerable care was put into the creation of the trial's golden list, which then informed the creation of the final golden list. A good idea of the problems to be dealt with was obtained from the trial – see a first overview in Santos, Costa & Rocha (2003).

Instructions for the compilation of the final golden list tried to simplify matters to some extent. First, all grammatical words were simply classified as GRAM; second, multiword expressions were banned from the golden list (although compared and taken into account in other contexts), and tokenization of verbs with enclitics and contractions, as one token only, was forced upon the systems (by postprocessing programs furnished by the organization). Several other decisions on what to include or exclude, of different arbitrariness degrees, will be presented in what follows.

Number of analyses

There is a subtle difference between a form having both a masculine and a feminine interpretation and being vague (marked "I") regarding gender (the same applies to other attributes; we use gender here for concreteness's sake). Most systems did not make this distinction, though. However, this has important consequences as to how to interpret a system's output and how to compare it to the golden list. Are we comparing two analyses or one? (The number is relevant since the success or failure of a system is a function of the number of the right (or wrong) analyses.)

After a long debate, the organization finally decided to use "I" consistently for gender and number (and thus words like the noun *ourives* and the adjective *antimotim* had only one entry, corresponding to four combinations of gender and number).3

Is capitalization part of morphology?

Does a morphological analyser also include a module with information about proper names? Or does this belong to semantics or to "named entity recognition"? Should it deal with properly written words only, or rather normalize its input? This is no longer morphology proper, but there is hardly a system that does not deal with (and rely on) this information. In fact, all participating morphological analysers relied more or less heavily in their own tokenization principles – which, far from being similar, made the organization's attempt (later dropped) to independently tokenize the texts harmful for all systems.

Our decision – also given the fact that the human compilers were often at a loss about the gender of e.g. city names, such as *Chaves* in Figure 1 – was to include in the golden list only number and gender for first names, such as *Ana*; and consequently not compare the values of these features for proper names. Surnames were not considered additional proper names, given that in principle any common noun can be used as a surname4; toponyms were included, but without gender.

Is derivation part of morphology?

[2] It should be mentioned that *Morfolimpíadas* also included a comparison based on running text (concentrated on tokenization), already covering frequency and coverage issues.

[3] Note that these decisions apply also to how the systems' output is processed, which must match the golden list assumptions. The organization had to cater for the corresponding encoding.

[4] An already identified proper name did not get an additional entry; but a capitalized word in the golden list, not known to be a first name, country name, etc. was marked as PROP.

Although this is obviously so from a theoretical linguistics perspective, the outcome of a derivation analysis was so far from consensual from the point of view of the computational systems involved, that we actually dropped it in the end. (In fact, only some systems had this capability.)

There was no agreement in the way to specify the result and the form of derivation, and even without considering the encoding of the derivational information itself, this caused problems in lemma comparison. In order to handle fairly systems with either philosophy, we were forced to state alternative analyses in the golden list, for all forms for which it would be possible to assign a derivational analysis.

```
encantadora÷ADJ÷encantar÷(...)÷deriv dor
encantadora÷ADJ÷encantador÷(...).÷.
encantadora÷SUB÷encantar÷(...)÷deriv dor
encantadora÷SUB÷encantadora÷(...).÷.
```

Figure 2: Double analysis of possible derivations

Now, both sides can adduce arguments for using a specific strategy. Thus, in Figure 2, the proponents of the first strategy would claim that one could be interested in finding all concepts related to the word *encantar* ("to charm"), while the defendants of the second strategy would claim that, linguistically, a noun or an adjective cannot have as the lemma (or base form) a verb, and that often a derivation results in different senses and different syntactic behaviour.

The problem for the organizers was that there was no way to automatically convert one analysis to another, since different strategies conveyed (and missed) different pieces of information.

While it was decided not to enter with derivation information in the last evaluation measures –because also the cases marked as derived varied widely across systems; some only marked derivation if the word had been heuristically derived by the system, i.e., was not in the lexicon, while others always marked it – in the (prior) process of creating the golden list, this had to be taken into consideration, and led in fact to the inclusion of "difficult" cases for both approaches, as Figure 3 shows.

```
retratado÷ADJ÷retratado÷.÷S÷.÷M÷.÷.÷.
retratado÷ADJ÷tratado÷.÷S÷.÷M÷.÷.÷deriv re
retratado÷V÷retratar÷PP÷S÷.÷M÷.÷.÷.
retratado÷V÷tratar÷PP÷S÷.÷M÷.÷.÷deriv re

inalterada÷ADJ÷inalterado÷.÷S÷.÷F÷.÷.÷.
inalterada÷ADJ÷alterado÷.÷S÷.÷F÷.÷.÷deriv
in
inalterada÷V÷alterar÷PP÷S÷.÷F÷.÷.÷deriv in
```

Figure 3: Non trivial cases for and against derivation

So, the form *retratado* can have two interpretations, corresponding to two separate meanings: one as the result of a proper derivation (with the prefix *re-*), and another related to the full lexical item *retratar* ("to portray"). Contrary to the case of Figure 2, one would plausibly want to have both analyses.

Conversely, while apparently the form *inalterada* looks like a past participle, there is no original verb **inalterar*, and it requires systems that encode derivation to give a sensible analysis for *inalterada* as a verb form: one must find the verb *alterar* as the origin of such form with the prefix *in-*. Similar cases are the word *fauvismo* (originating from French *fauvisme*, from French *fauve*, which is not a Portuguese word) and *belissimamente*.

The infamous problem of past participles

One other complex problem – already foreseen at the outset – had to do with past participles and adjectives. Instead of describing the situation (for this see e.g. Barreiro, 1998), we want to show here the problems it brings to a fair comparison of systems, and the way we tried to deal with it in the golden list.

The participating systems reflected three[5] different ways of handling forms in *-ada*, *-ados* and *-adas*: considering them only past participles of a verb, considering them only adjectives, and considering them both. Only the third alternative would allow for differentiation of cases where only one of the two interpretations might be possible, as is the case of *resultado*[6] (verbal) and *inalterada* (adjectival). Still, to choose the third option as the correct one was to consistently penalize the systems with options one or two, and this goes against the cooperative philosophy of the whole endeavour.

So we had to consistently include the adjectival and verbal interpretations of these forms in the golden list, but use a complex evaluation function that only took into account either adjective or verbal reading for designated systems. The most interesting outcome from our perspective, in fact, was to find out that this was a parameter that should be taken into account when a user chose a morphological analyser for Portuguese, and that there was a tripartite division in the approach to the analysis of these forms.

Part of speech unresolved issues: noun or adjective?

Already in Medeiros, Marques & Santos (1993), it had been suggested to use, for morphology, the category n/a (noun/adjective) for many Portuguese words (and leave to the syntactic context the decision of assigning a nominal or adjectival role in most cases[7]). To be sure, also at the syntactic plane there are many levels where a particular piece of information can be encoded, with or without change in the actual assignment of PoS category of a given word, as illustrated in Santos & Gasperin (2002). Still, and given that all systems involved encoded both adjective and noun readings of a given form, the golden list creation followed this traditional path.

This choice was not without problems, though, especially for feminine adjectives. While in the case of Figure 2 it was straightforward to assign both adjective and noun readings to *encantadora*, given that there is a role (that of snake incantator) for which there are no

[5] Two, in case of forms in *-ado*. No system considers them only adjectives.

[6] The adjective corresponding to the verb *resultar* is *resultante*.

[7] In fact, not every case requires – or has – a consensual analysis, even in context. Santos (2003) reports cases where individual PoS are problematic, but not the classification of a higher level constituent. On the other hand, in a lexicon with 35,000 noun/adjective candidates (Barreiro, Pereira & Santos, 1993), only 5,000 items had both noun and adjective readings.

gender constraints; cases like *feliz* ("happy") and *bravo* ("braveheart"), prototypical adjectives who have also (masculine) noun entries in a published dictionary – respectively a kind of clown and a courageous person – were more difficult to classify. We decided not to add seemingly dubious far-fetched readings, if they were not also explicitly dictionarised as feminine nouns. (Note that regular forms in *-ora* are missing also from most printed dictionaries8 when simply the feminine counterpart of nouns in *-or*; and that any adjective can appear isolated in Portuguese with an anaphoric purpose – an expletive pronoun like "one" in English is not required.)

The question of rare (morphological) readings

Even though we did not perform a frequency analysis to decide which forms should appear in the golden list, we found the need to indicate cases where there were central uses (and interpretations) and very peripheral ones (often unknown to us, but found after Web search or in some published paper dictionaries). The marking was not difficult (although it is important to note that "rare" means relative to the frequency or centrality of the other readings, not "absolute" rare; cf. *tinha* as disease name vs. verb form of *ter*, "have"). We thought that systems would differ considerably in the way they would perform precisely on these rare items, but found that, on the contrary, every system performed better if the rare interpretations were left in the golden list.

The result

In Table 1 we present a quantitative overview of the golden list. The distribution of the forms in the texts, as well as the texts themselves, can be found in the website.

Kind	Size	%
Forms	655	
Hyphenated forms	30	4.6%
Verbs with clitics	16	2.4%
Deviating forms	19	2.9%
Forms with SUB analysis	409	62.4%
Forms with V analysis	297	45.3%
Forms with ADJ analysis	257	39.2%
Forms with PROP analysis	44	6.7%
Forms with one analysis	210	32.1%
Forms with two analyses	236	36.0%
Forms with three analyses	96	14.7%
Forms with four analyses	80	12.2%
Forms with more than four analyses	33	5.0%

Table 1: The golden list at a glance

Concluding remarks

This is a tiny subset of the interesting issues that arose in the preparation of Morfolimpíadas. Some of them are mainly interesting for the Portuguese language processing community, others – we hope – are of general interest to document the concerns associated with evaluation contest paradigm.

One of the most striking conclusions was that there is still a lot of both theoretical and practical disagreement in an apparently "simple" task, concerning how to handle a myriad of problems a morphological analyser for real text has to deal with. In fact, a practical task definition brings about many relevant practical problems that do not concern theoreticians.

Acknowledgements

First of all, we are very grateful to the *Morfolimpíadas* participants. This paper describes the decisions taken by the organizing committee as a whole, not by the authors. We thank the other golden list compilers: Susana Afonso, Luís Costa and Paulo Rocha, as well as all members of the scientific committee, who participated in the golden list compilation. The work was supported by Fundação para a Ciência e Tecnologia through grant POSI/PLP/43931/2001, and co-financed by POSI.

References

Barreiro, Anabela M. (1998). Propriedades Sintáctico-Semânticas dos Particípios Passados em Português Europeu. MSc thesis, Universidade Nova de Lisboa.

Barreiro, Anabela, Maria de Jesus Pereira & Diana Santos (1993). Critérios e opções linguísticas no desenvolvimento do Palavroso, um sistema computacional de descrição morfológica do português. INESC Report RT/54-93, December 1993.

Hausser, Roland (ed.) (1996). Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994. Tübingen: Max Niemeyer Verlag.

Hirschman, Lynette (1998). The evolution of Evaluation: Lessons from the Message Understanding Conferences. Computer Speech and Language 12 (4), 281--305.

Medeiros, José Carlos, Rui Marques & Diana Santos (1993). Português Quantitativo. In Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93 (pp.33--38). Lisboa.

Santos, Diana (2003). Timber! Issues in treebank building and use. In N. J. Mamede, J. Baptista, I. Trancoso & M.G.V. Nunes (eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings (pp. 151--158). Berlin: Springer Verlag.

Santos, Diana, Luís Costa & Paulo Rocha (2003). Cooperatively evaluating Portuguese morphology. In N. J. Mamede, J. Baptista, I. Trancoso & M.G.V. Nunes (eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings (pp. 259--266). Berlin: Springer Verlag.

Santos, Diana & Caroline Gasperin (2002). Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In M.G. Rodríguez & C.P.S. Araujo (eds.), Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation (pp. 597--604). ELRA.

Santos, Diana & Paulo Rocha (2003). AvalON: uma iniciativa de avaliação conjunta para o português. In A. Mendes & T. Freitas (orgs.), Actas do XVIII Encontro da Associação Portuguesa de Linguística (pp. 693--704). Lisboa: APL.

[8] The issue of comparing and assessing the information in published dictionaries would require an article on its own; let us simply note that we used six different lexicographic sources, Web search, and our intuitions, and that the trickier decisions are documented in the website.