Cooperatively evaluating Portuguese morphology

Diana Santos[1], Luís Costa[1], and Paulo Rocha[2]

[1] Linguateca, SINTEF Telecom & Informatics
Pb 1124 Blindern, 0314 Oslo, Norway
{Diana.Santos,Luis.Costa}@sintef.no
www.linguateca.pt
[2] Linguateca, Departamento de Informática, Universidade do Minho
Campus de Gualtar, 4710-057 Braga, Portugal
Paulo.Rocha@di.uminho.pt

Abstract. This paper describes the first attempt to evaluate morphological analysers for Portuguese with an evaluation contest. It emphasizes the options that had to be taken and that highlight considerable disagreement among the participating groups. It describes the trial intended to prepare the real contest in June 2003, its goals and preliminary results.

1 Introduction

Morphological analysers for Portuguese constitute an almost unavoidable first step in natural language processing. No matter whether one is mainly concerned with parsing, information retrieval or corpus exploration, most groups have to deal with some form of morphology.

It is therefore no surprise that, in the spirit of our project, the first evaluation contest organized was in the realm of morphology.[1] Not surprising, either, that we were able to run a trial with six different systems in September/October 2002, and are expecting more participants in the forthcoming Portuguese Morpholympics (the *Morfolimpíadas*) scheduled for June 2003.

The evaluation contest model is well known in the NLP community and we refer elsewhere to its description [1,2]. As far as we know, it has only once been applied to morphology for German, the 1994 Morpholympics [3].

Inspired by the message understanding conferences (MUC) model, we performed an initial trial to evaluate several possible courses of action and help organize the larger event later, taking into account Hausser's experience and recommendations.

The evaluation should be based both on a set of forms with the "right" analysis (the golden list), and on scores based on larger amounts of text, for which there was no previous solution. The trial should be as similar as possible to the final event, but the

[1] There was a preliminary hearing of the Portuguese (language) NLP community on which areas there was more interest in evaluating. Morphology, corpora and lexica came up first. See http://www.linguateca.pt/AvalConjunta/.

scores should not be made public. Trial participants should benefit not only from the experience but also in that they constitute the organising committee of the major event and help shaping its final form.

In addition to run a rehearsal of the competition, we wanted to a) try out which input and output form was better; b) test whether it was possible to create a golden list which was representative of morphological knowledge required and of morphological problems to be solved; c) investigate whether there were significant performance differences per text kind, variant, genre, etc., and d) find measures that could adequately represent performance and highlight meaningful differences between systems.

2 Test materials creation

First, we asked the participants to cooperatively build a golden standard, by sending us a set of 20-30 judiciously chosen forms with the right analysis to be used in the contest, preferably in the format of their system. It was stressed that the analysis of those forms did not need to reflect real performance; rather, it should represent ideal performance. The first task was to put together the different items sent by 8 different sources for inclusion on the golden list, while at the same time compiling a set of test texts that wrapped all forms, preventing the golden list items to be identified.

2.1 Test texts

The test texts were amassed by randomly extracting chunks including the forms in the golden list from a wide variety of distinct sources, maximizing the set of different available categories, such as subject area, kind of newspaper, translated text or not, etc, over all corpora available (see table 1).

Table 1. Distribution of the test texts (according to the organization's tokenization, reflected in the **uul** format)

Variant	Words	Texts	Genre	Words	Texts
Total	39,850	199	Total	39,850	199
Brazilian	16,132	82	Newspapers	23,823	118
Portuguese	21,206	113	Original fiction	836	3
African	1,390	4	Translated fiction	3,117	18
Unknown	512	1	Web / email	3,333	19

The texts were provided to the participants in three formats: **uts**, an alphabetical list of the (single word) types in all texts, accompanied by a small **mweuts,** an alphabetical list of a few possible[2] multiword types in the texts; **uul**, a pre-tokenized text, one word per line; and **ts**, running text.

The participants had one week to provide the output of their morphological analysers of all forms in sequence, for each of the four files. (A morphological disambiguation task was also included for those groups having systems that could do it. In that case, the goal was to find the right form in context.)

The output of the systems should then be compared to the right answers in the golden list; and compared in general among the systems. This pressuposed the translation of each output into a common format, and the creation of programs for comparison and evaluating the results. The whole organization philosophy gives the least work to the participants, putting the burden on the organization.

2.2 Golden list compilation

Before any of these tasks could be undertaken, however, it was necessary to produce a golden list on which all participants agreed. Maybe not surprisingly, the compilation of the golden list turned out to be an extremely complex task, and involved an endless number of versions until its final form.

Let us give here some simple statistics: the final golden list contained 200 forms having 345 analyses (on average, 1.73 analyses per form). 113 forms had one analysis, 52 two, 20 three and 15 four or more. 114 analyses pertained to verbs, 95 to nouns, and 14 were labelled proper nouns. Defining weight as the ratio of (e.g., verb) analyses over all analyses of the forms which have verb readings, we find verb weight, noun weight and proper noun weight to be .61, .540 and .583 respectively. Of the entries, 9 were multiword expressions, 5 were cliticized verbs and 7 contractions, 14 were hyphenated (different kinds) and 3 forms were deviant (with one analysis only), including foreign words, common spelling mistakes, and neologisms.

Even though the trial organizers had restricted the competition to quite consensual categories (we thought) such as gender, number, lemma or base form, tense, as well as occurrence of superlative or diminutive, we found out that there was a surprisingly larger span of fundamental differences between systems than expected.[3] By preliminary inspection of the participants' output formats for a set of forms, and using common knowledge about Portuguese, we had already come up with a set of encoding principles, in order to minimize encoding differences, e.g.:

- whenever there is total overlapping between verbal forms, we encode only one (as is the case of 3rd person singular personal infinite and impersonal infinite; 3rd person plural Perfeito and Mais-que-perfeito; etc.)
- we joined as one form verbs with clitics and erased as irrelevant (for purposes of common evaluation) all sublexical classifications: e.g. for *fá-lo-ia* ('would do it'), some systems would return separately all possibilities for *fá*, for *lo* and for *ia* as

---

[2] Containing both some commonly considered idioms or locutions, as well as sequences of two or three words that just occurred in sequence.

[3] The categories were extensionally defined as the pairwise intersection of what was provided by the actual systems, a sample of which output having been requested at an early stage.

independent words.

Still, while looking at the set of golden candidates chosen by each participant we had to confront much deeper disagreement, as listed in the next section.

2.2.1 Different linguistic points of view

During the process of harmonizing the golden list, we found the following "theoretical disagreement" (as opposed to actual occurrence of different analyses of specific items): differences about PoS categorization (cases 1-4); differences about which information should be associated with a given PoS (5-7); differences about base form or lemma (these are perhaps the most drastic and the ones that affect the larger number of golden list items) (8-10); differences about the values of a given category (11-12); and differences on what should be done by a morphological analyser (13-14).

1. some researchers would have a given word ambiguous between noun and adjective; others considered it belonged to a vague noun-adjective category
2. some words were either considered proper names or nouns
3. quite a lot or words were either considered past participle or adjective or both
4. there was no agreement on whether a morphological analyser should return the PoS "adverb" for *clara*, given that in some contexts, adverbs in *mente* drop the *mente* suffix, as in *clara e sucintamente*
5. some systems do not consider gender as a feature of past participles
6. gender/number for proper names: there is internal gender but also a proper name can be used often to identify all kinds of entities
7. should gender be assigned to pronouns when they are invariable?
8. when adverbs in *mente* are related to adjectives, some systems return the adjective as lemma, others the adverb form
9. derived words: the systems that analyse derivation are obviously different in kind and in information returned from those that only handle inflection, but it seems that there is no standard encoding of derivation information, either
10. for some hyphenated words, that have more than one "head", there seems not to be a consensus about how to represent their lemma(s)
11. is indeterminate a third value for gender, or it means both M and F?
12. how many tenses are there, are tense and mood different categories?
13. are abbreviations and acronyms in the realm of morphology?
14. should a morphological analyser return "capitalized", "upper case", "mixed case" and so on as part of its output? This question arose because not all morphological analysers actually return the input form, so this information may be lost.

Then, we have to mention the well known hard problems in morphological processing: differences in MWE handling; differences in clitic and contraction handling, and differences in the classification of closed words. (In fact, the organization had initially stated that the classification of purely grammatical items was not interesting for an evaluation contest, since they could be listed once and for all as they concern a small number of words, but we still had to deal with them due to forms which had both grammatical and lexical interpretations.)

2.2.2 Absence of standard

The cooperative compilation allowed also a general acknowledgement that there was no standard available for specially formatted areas (such as bibliographic citations, results of football matches, references to laws); traditional spelling errors; foreign words; oral transcription; and random spelling errors.

It was agreed that for these cases one should simply count their number in the texts, not use them for evaluation. However, this is obviously easier said than done, since they cannot be automatically identified.

2.2.3 Different testing points of view

The compilation process also showed that there were different views of what a golden standard should include, from really "controversial" items, to multiword expressions and punctuation (none of these had been foreseen by the organisers, although later accomodated).

Some participants thought it would be enough to give one analysis (the one they wanted to test) and not all analyses of one form; some were intent on checking the coverage of particularly ambiguous forms; others to see whether a rule-generated system would block a seemingly regular rule to apply.

Some of these differences could and should be solved by clear compilation guidelines and neat examples – like "give all analyses of the form you selected"; "do not include purely grammatical items in the golden list candidates"; "comment or mark deviant items", etc., but others are really interesting since they reflect genuinely different system conceptions with correspondingly different ways of testing them.

Finally, it should be mentioned that, although the problem of rare forms and their possible relevance in the context of an evaluation of morphological analysers was debated, no solution has yet emerged.

3 Measuring

The next step of the trial was to process each system's three outputs and translate them into an internal evaluation format. Then, another set of data was gathered by extracting the data relative to the elements present in the golden list for subsequent comparison of their classifications.

3.1 Tokenization data

The rationale behind the three formats was our fear that too many tokenization differences would hinder easy comparison of running text results ([4] reports 12-14% tokenization differences between two different systems for Portuguese). We provided the **uul** format in order to provide a common tokenization, but this did only half the job, anyway, since some systems still separated our "units" into smaller parts, while others joined several of them. On the other hand, while the **uts** format prevented joining into longer units (since it is an alphabetical order of all types in **uul**), it was unrealistic or even unfair in that it might provide the systems with tokens they wouldn't find if they were left in charge of tokenization.

By providing the same texts in three forms, we wanted to check how significant would be the differences, and eventually choose which was best, or whether a combination (probably measuring different things in different formats) should be used.

Quantitative data, in terms of coverage and tokenization agreement, are displayed in tables 2 and 3. Tokenization differences imply that, even if all systems returned exactly the same analyses for the forms they agreed upon, there would still be disagreement for 15.9% of the tokens, or 9.5% of the types.

Table 2. Tokenization overview, for the **ts** format. A token is considered common if it was found by the four systems. 8480 tokens were common. One token can have several analyses

System	B	C	D	E
No. of tokens	41,636	41,433	39,503	41,197
No. of analyses	73,252	76,455	57,650	69,619
Common tokens	84.1%	91.6%	86.5%	86.2%

Table 3. Tokenization overview, for the **uts** format (one type only). A type is considered common if it was found by the four systems. 9580 types were common

System	B	C	D	E
No. of types	11,593	10,896	10,613	10,745
No. of analyses	18,483	18,742	15,005	13,487
Common types	90.7%	92.0%	91.3%	90.5%

As to whether the performance of the morphological analysers varied significantly with text kind, table 4 shows some first results, concerning the analysers' performance regarding language variant (BP: Brazilian, PP: from Portugal).[4] (Internal) coverage is defined as the percentage of the tokens identified by the system for which it could produce some analysis (as opposed to "unknown").

Table 4. Impact of variant in internal coverage and number of analyses, for the **ts** format, after handling clitics and contractions, and counting every (set of) grammatical analysis as one

System	B	C	D	E
Coverage PP	98.20%	99.95%	99.19%	94.94%
Analyses/form PP	1.38	1.67	1.26	1.44
Coverage BP	97.20%	99.85%	98.46%	96.40%
Analyses/form BP	1.38	1.65	1.26	1.42
Total coverage	97.58%	99.87%	98.82%	95.65%
Analyses/form gen.	1.39	1.62	1.26	1.43

[4] There are well-known morphological and (regular) ortographical differences between the two variants. However, it is possible that most systems can cope with those. On the other hand, results can always be parametrized, for variant-specific systems.

3.2 Comparison with the golden list

For comparison with the golden list, one can use two different units: the analysis and the form, which may in turn have several analyses. In table 5, not including the two punctuation items, we first give the number of forms and analyses provided by each system, and then proceed to count forms whose set of analyses is exactly like the golden list (column 3), individual analyses just like the golden list (column 4), and which forms had an altogether different number of analyses (column 5), subdivided in more analyses and less analyses. Finally, we look at the set of PoS for a given form (column 8). Later, we intend to use "meaningful combinations of features" [5]. It can in any case be reported that the highest number of differences concerned lemma.

Table 5. System comparison with the golden list, using **uts**. Each *form* (ambiguation class) was compared, as regards number of analyses, and set of PoS classifications assigned

System	no. forms	no. analyses	equal forms	equal anal.	diff. no.	more	less	diff. in PoS set
golden li	198	343	198	343	0	0	0	0
system B	168	297	26	101	69	32	37	70
system C	178	315	93	192	50	24	26	45
system D	186	299	64	160	59	23	36	72
system E	182	274	83	145	67	14	53	83

Note that the column for "no. forms" is necessary since not all forms were recognized as such by the competing systems (in addition, the **mweuts** data are not included for lack of an easy comparable unit). Several scoring functions can then be applied.

3.3 Coarse-grained comparison of the output for all tokens

We wanted to see whether it was possible to measure (blind) agreement among systems based on all common tokens recognized. Table 6 presents an initial “agreement table”, where the system in the left is considered as correct and the system on top is measured against it, after a first homogeneization procedure, concerning contractions and clitics, was applied, and all grammatical analyses were reduced to one.

Table 6. System cross-comparison, based on **uts**. Each system in turn is used as golden standard. The cells contain the percentage of analyses of the top system which agree, of all analyses of the system on the left). The field "other information" was not taken into account

System	B	C	D	E
system B	100%	68%	51%	47%
system C	69%	100%	57%	53%
system D	66%	68%	100%	53%
system E	64%	71%	59%	100%

These are the raw numbers. Although they may seem overwhelming, several steps can be taken to reduce sytematically some of the differences, not only by harmonizing further the tokenization (such as numbers, proper names and abbreviations), but especially by taking into consideration systematic conflicting points of view.

In fact, thanks to the discussions among the participants on what should or should not be considered right, it was possible to trace several cases of theoretical disagreement, where one might define not one but several evaluation (and thus, comparison) functions depending on the theoretical standpoint.

After giving the question thorough consideration, we decided to compute a minimum information function (*minif*) and a maximum information function (*maxif*), and make our comparisons and evaluations based on these. In practice, this means to define two internal evaluation formats, and use more complex translation procedures.

Finally, there are several other evaluation methodologies investigated, which for lack of space cannot be presented here, and will hopefully be published elsewhere: comparison with already annotated corpora; comparison with the (automatic) output of the disambiguation track; finer measures of which lexicon items are more prone to disagreement, and so on.

We are also investigating the semi-automatic compilation of a new golden list, dubbed the *silver list*, following a proposal at the trial meeting in Porto.

This paper should, in any case, have enough material to give a glimpse both of the complexity of the organization of the forthcoming Morfolimpíadas (see http://www.linguateca.pt/Morfolimpiadas/ for updated information) and of the many issues in the computational morphology of Portuguese when applied to real text.

It should be emphasized that this paper (and the contest whose first results are here presented) is only possible through the cooperation (and effort) of many people. We are greatly indebted to all participants in the trial, as well as to the researchers and developers who have contributed with suggestions, discussion or criticism.

References

1. Hirschman, Lynette: The evolution of Evaluation: Lessons from the Message Understanding Conferences. Computer Speech and Language **12** (1998) 281—305
2. Santos, Diana, Rocha, Paulo: AvalON: uma iniciativa de avaliação conjunta para o português. In: Actas do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002) (2003)
3. Hausser, Roland (ed.): Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994. Tübingen: Max Niemeyer Verlag (1996)
4. Santos, Diana, Bick, Eckhard: Providing Internet access to Portuguese corpora: the AC/DC project. In Gavriladou et al. (eds.): Proceedings of LREC'2000 (2000) 205-210
5. Santos, Diana, Gasperin, Caroline (2002). Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In Rodríguez, M.G. & Araujo, C.P.S. (eds.): Proceedings of LREC 2002 (2002) 597-604