An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese*

Sandra Aluísio[1,2], Jorge Pelizzoni[2], Ana Raquel Marchi[2], Lucélia de Oliveira[2], Regiana Manenti[2], and Vanessa Marquiafável[2]

[1]ICMC – DCCE, University of São Paulo, CP 668, 13560-970 São Carlos, SP, Brazil
{sandra,jorgemp}@icmc.usp.br
[2]Núcleo Interinstitucional de Lingüística Computacional (NILC), ICMC-USP, CP 668
13560-970 São Carlos, SP, Brazil
{raquel,lucelia,regiana,vanessam}@nilc.icmc.usp.br

Abstract. This article identifies and addresses the major linguistic/conceptual, as opposed to logistic, issues faced in the morphosyntactic tagging of MAC-Morpho, a 1.1 million word Brazilian Portuguese corpus of newspaper articles that has been developed in the Lacio-Web Project. Rather than simply presenting the annotated corpus and describing its tagset, we elaborate on the criteria for establishing the tagset and analyze some interesting cases amongst the linguistic problems we faced in this work.

1 Introduction

Annotated reference corpora, such as Suzanne, the Penn Treebank or the BNC have helped both the development of English computational linguistics tools and English corpus linguistics. Manually-annotated corpora with part-of-speech (POS) and syntactic annotation are costly but allow one to build and improve sizeable linguistic resources, such as lexicons or grammars, and also to develop and evaluate most computational analyzers. Usually, such treebank projects follow the Penn Treebank (http://www.ldc.upenn.edu/Catalog/docs/treebank2/cl93.html) approach, which distinguishes a POS tagging and a parsing phase each comprising an automatic annotation step followed by manual revision. Recently, there have been several efforts to build gold standard annotated corpora for other languages than English, such as French, German, Italian, Spanish, Slavic (http://treebank.linguist.jussieu.fr). For Brazilian Portuguese (BP), however, the figure is not so bright. With regard to manual morphosyntactic annotation, to the best of our knowledge, there are only two small Brazilian corpora which were used to train statistical taggers: (i) the 20,982-word Radiobras Corpus [1, 2], and (ii) the 104.966-word corpus built from NILC's corrected text base spanning 3 genres (news, literature and textbooks) [3]. There are, although, several (Brazilian and European) Portuguese corpora automatically annotated by Bick´s [4] syntactic parser PALAVRAS (http://visl.hum.sdu.dk), which are part of the AC/DC project (http://www.linguateca.pt).

* This project is partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We are grateful to E. Bick for parsing MAC-Morpho.

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 110–117, 2003.

In order to make freely available both corpora and computational linguistic tools which learn from raw and annotated corpora, such as POS taggers, parsers and term extractors, we have started the Lacio-Web project. Lacio-Web (LW), a two-year project launched at the beginning of 2002, tries to fill the gap with regard to linguistic resources and tools for BP. In this paper we present the rationale for building a 1.1 million-word corpus with manually validated morphosyntactic annotation (the results of the inter-annotator agreement evaluation and further logistic/historical detail have been published in [5]), including the criteria for establishing the tagset (Section 2), some linguistic problems we faced in this work (Section 3) and directions for further work (Section 4). This corpus was taken from a text collection from Folha de São Paulo (http://www.folha.uol.com.br/folha), which gives us high quality contemporary Brazilian Portuguese from different authors and domains. The resulting annotated corpus (named MAC-Morpho) will be available in two versions: in annotators' format (one word per line followed by its tag) and in the XML-compliant format proposed by the EAGLES [6] (www.cs.vassar.edu/XCES).

2 Designing the Tagset

We analyzed the Eagles recommendations for the Morphosyntactic Annotation of Corpora (http://www.ilc.pi.cnr.it) and two of the more important tagsets designed for English (36-tag Penn Treebank Tagset and BNC project's[1] 61-tag C5 and 140-tag C7) and three other tagsets for Portuguese (NILC[2], PALAVRAS and Tycho Brahe [7] respectively with 36, 14 and 48 tags). Although there are already two tagsets for Portuguese (PALAVRAS and NILC), whose purpose is similar to ours, neither fulfills all the criteria we consider as essential to our project. These criteria have been employed by and large by the Penn Treebank and Tycho Brahe projects. Even though the latter project also tackles Portuguese, it has been specifically designed to support diachronic research and, perhaps due to this, ends up with a conceptually different tagset from ours.

2.1 Criteria, Features, and Previous Work

Recoverability. Exploiting recoverability refers to avoiding tagging (morphological) details that can otherwise be easily recovered by querying a lexicon on the basis of the word and its tag alone. For example, the decision of having a unified "article" tag – instead of two or more, such as "definite/indefinite singular masculine article" – takes advantage of the automatic recoverability of any further features of interest, provided articles are not ambiguous with each other. This criterion ultimately leads to minimal tagsets with the sole purpose of disambiguation, i.e., a tagset suffices as long as every possible pair (word, tag) resolves to at most one single lexical entry (whatever an entry may be) or set of morphologically equivalent entries. NILC Tagset fails to exploit, for instance, the recoverability of the traditional Portuguese pronoun classes,

[1] http://www.hcu.ox.ac.uk/BNC/what/ucrel.html
[2] http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm

ending up with 10 distinct pronoun tags. Were we to satisfy recoverability solely, 2 simple tags ("relative and non-relative pronoun") would do exactly the same effect.

Syntactic Function (and Actuality). Notwithstanding, recoverability and its related morphological disambiguation efficiency are not enough, since we strictly understand that the ideal tagset should be optimal for supporting a subsequent full syntactic parsing step. In other words, it should entail as much syntactical inference as possible while not requiring its tagger to be a full-fledged parser, paradoxical though it may seem. Thus, recoverability is but a lower-bound measure, ever second to syntactic function, an eminently tag-multiplying factor.

The referred paradox is not trivial, and the pitfall of reaching a fully syntactic, or simply overcrowded, tagset may seem unavoidable, at first sight. Fortunately, we believe we quite managed to develop a twofold sound compromising criterion, namely:

- *intra-word syntactic Distinctness preservation* (or D-preservation): any two syntactically distinct occurrences of a word should never receive the same tag;
- *inter-word syntactic Likeness preservation* (or L-preservation): reciprocally, any two syntactically equal occurrences of different words should receive the same tag as long as morphological recoverability is left unharmed.

The application of D-preservation to our former two-tag treatment of pronouns ("relative" vs. "non-relative") leads to LW Tagset's five pronoun tags, namely PROPESS (personal pronoun, of whatever grammatical case), PRO_KS_REL (relative subordinating pronoun), PRO_KS (non-relative subordinating pronoun, introducing noun clauses, such as "who" in "Please identify who the murderer is."), PROSUB (non-subordinating, non-personal pronoun as a nucleus, such as "who/this" in "Who/This is the murderer?") and PROADJ (non-subordinating, non-personal pronoun as a modifier, such as "this" in "This man is the murderer."). In these examples and in accordance with the stated criterion, two syntactically distinct occurrences of "who/this" receive accordingly distinct tags. It is worth noticing that, properly exploiting recoverability and syntactic encoding, our five-tag treatment of pronouns is more informative than that of NILC Tagset, despite the latter having twice as many pronoun tags.

In time, syntactic function implies syntactic actuality, i.e., tags should clearly reflect the syntactic function of words in the clauses and phrases they belong to, which sometimes means departing from traditional (usually untenable) treatment. One such example is the introduction of the tag ADV_KS_REL (relative subordinating adverb) to account for relative "(P) onde // (En) where", "quando // when" and "como // how" (the latter is never relative in English, but arguably so in Portuguese), traditionally regarded as pronouns. That is not an unheard-of position, since PALAVRAS also treats these words as adverbs. But maybe a bit too eagerly: according to its POS tagset, e.g. "quando // when" is always an adverb, whereas we understand it may fall into four categories, namely KS (subordinating conjunction, in adverbial clauses), ADV_KS_REL (relative subordinating adverb, in relative clauses), ADV-KS (non-relative subordinating adverb, e.g. in indirect interrogative sentences) and plain ADV (non-subordinating adverb, e.g. in direct interrogative sentences). To do PALAVRAS justice, however, we should notice that it is a parsing system, not a POS tagger, and its performance seems to be not at all hindered by such simplifications, which is the case exactly because (i) it is not based on the more common tagger-parser pipeline architecture and (ii) it avails itself of a host of secondary morphosyntactic tags. The application of L-preservation is exemplified while discussing the immediately following criteria.

Consistency and Indeterminacy. A tagset is worth nothing if it does not provide for consistency, i.e. if its users (not only corpus annotators) are not likely to agree (including with themselves!) on how and when to use each tag. Even if we only employed one single all-consistent, all-efficient annotator, users must be able to evaluate, understand and ultimately replicate their work. The pursuit of consistency is paramount, even if to the detriment of other requirements. In specific, consistency is not usually very partial to refinement, which here means syntactic or morphological detail. One such example is the contrast between past participles in adjectival position (e.g. "(P) a casa pintada // (En) the house (that has been) painted") and adjectives proper zero-derived from past participles (e.g. "(PBr) uma moça muito falada // (En) a young woman very much gossiped[3] about"), whose annotation was intended by the Lacio-Web team at first, but had to be eventually abandoned due to low inter-annotator consistency. The solution here was to resort to indeterminacy, introducing the (indeterminate) PCP tag, standing for "past participle or adjective zero-derived therefrom". Indeterminate tags are created by collapsing inconsistency-mongering tags, thus leading to smaller tagsets.

Nonetheless, it is not always that indeterminate tags are the best solution for inconsistency problems. Sometimes, just sound application of other criteria might come to one's rescue. One ever-lasting source of debate and inconsistency in Portuguese has been the contrast between nouns and adjectives. Unlike their English counterparts, most Portuguese nouns and adjectives can be used interchangeably, making it hard to determine the actual morphological specification of these words and whether nominalization is really taking place, so used to this operation are we native speakers. By simply prioritizing syntactic function, or rather, by upholding L-preservation, we were able to circumvent this delicate problem, the result being thus: every open-/closed-class occurrence happening to be the nucleus of a noun phrase is tagged N/PROSUB; and every open-/closed-class occurrence happening to modify a noun, ADJ/PROADJ or ART (article, whether definite or not). Even the words traditionally called "numerals" usually fall into either N or ADJ, again according to the syntactic function of each occurrence. Only cardinal numerals and all inflections of the word "(P) meio // (En) half" may receive the tag NUM (numeral), and do so only when occurring as noun modifiers, due to their remarkably distinct syntactic behavior in such cases. Therefore, those "numerals" never happening to be real noun modifiers (e.g. "bilhão/milhão // billion/million", "dezena // ten", "terço // third", "quarto // quarter") will never be tagged NUM.

Learnability. Finally, we cannot fail to mention that a most limiting factor to how syntactic LW Tagset could get was, at all times, the assumption of a machine learning technology to apply to (a version of) the annotated corpus, namely that usual in POS taggers and blind but to a very few words contiguously surrounding the current target word. Therefore, it seemed just fair to avoid all refinement that was really not likely to be learnt, such as NILC Tagset's annotation of verb transitivity.

[3] Notice that, unlike English "gossiped", Portuguese "falada" cannot be accounted for by productive passive voice processes. That is exactly why the latter is regarded as a zero-derived adjective proper.

It is worth noticing at this point that it has never been our aim to deliver a ready-to-use training corpus, but rather one providing for (i) rapid (i.e. automatic) deployment of variously tagged (e.g. for various levels of refinement) training versions of itself and thus (ii) extensive and comprehensive experimentation. Just by way of illustration of how not ready to use our corpus is, it should suffice to mention that some of its tokens are actually groupings of contiguous tokens in the original, resulting in what we call "compounds" (morphosyntactic units made up by two or more words, such as "(P) devido=a // (En) due=to"), which are tagged regularly as if they were but one single word. Rather more training-friendly, in contrast, NILC Tagset also employs multiword morphosyntactic units, but tags each of their tokens separately with the same tag. Naturally, contiguous multiword units having the same tag will pose a segmentation problem to NILC Tagset's users.

2.2 The Current Tagset

Since the beginning of its development, in July of 2002, LW Tagset (Tables 1 and 2) has undergone cyclic revisions, being currently in its ninth version.

Table 1. Regular tags

Tag	Definition
ADJ	open-class noun modifier
ADV-KS-REL	relative subordinating Adverb
ADV-KS	Non-relative subordinating Adverb
ADV	Non-subordinating adverb
ART	Article
KC	coordinating conjunction
KS	coordinating conjunction
IN	interjection
N	open-class noun phrase nucleus
NPROP	proper noun
NUM	numeral as a noun modifier
PCP	past participle or adjective
PDEN	emphasis/focus
PREP	preposition
PROPESS	personal pronoun
PRO-KS-REL	relative subordinating pronoun
PRO-KS	Non-relative subordinating pronoun
PROSUB	non-subordinating pronoun as a noun phrase nucleus
PROADJ	Non-subordinating pronoun as a modifier
VAUX	Auxiliary verb
V	Non-auxiliary verb
CUR	Currency symbol

Table 2. Complementary tags

<table>
<tr><th>Compl. Tag</th><th>Definition</th></tr>
<tr><td>|EST</td><td>foreign</td></tr>
<tr><td>|AP</td><td>apposition</td></tr>
<tr><td>|+</td><td>contraction/ enclitic</td></tr>
<tr><td>|!</td><td>mesoclitic</td></tr>
<tr><td colspan="2">| [beginning,
|... middle part,
|] and end of discontinuous compound (further discussed in Section 3)</td></tr>
<tr><td>|TEL</td><td>phone number</td></tr>
<tr><td>|DAT</td><td>date</td></tr>
<tr><td>|HOR</td><td>time</td></tr>
<tr><td>|DAD</td><td>formatted data not falling into above categories</td></tr>
</table>

At present it comprises 22 regular POS tags along with nine orthogonal complementary tags. The latter are thus called because they add to the information of the POS tags, to which they are optionally appended by means of the "|" symbol.

3. Some Emblematic Linguistic Challenges

NPROP – Proper Noun. In most respects, proper nouns are but nouns, especially in the relation they bear to noun phrases. What sets them apart is the prerogative to refer to one single entity of the real world in that, if X is a proper noun, X might even be shared by more than one entity (e.g. homonymous people), *but that would imply no common properties whatsoever to sharers*. Consequently, we should tag NPROP all those words that would otherwise be tagged N but happen to have strictly unitary extensions/indeterminate intensions. Such is our criterion for identifying proper nouns, which, clear though it may seem, makes plenty of room for inconsistency. Problematic cases usually fall into the following categories:

- **motivated NPROPs**, or rather, those obtained by zero-derivation, e.g. "(PBr) Nordeste (Brazilian geopolitical unit) // (En) the Northeast", "Congresso // the Congress";
- **metonymical NPROPs**, e.g. "(PBr) gillette // (En) (brand of) razor blade", "band-aid", "danone // (brand of) yogurt", "fusca // a specific make of car or car of this make";
- **NPROPs with context-dependent cardinality extensions**, e.g. "(P) sol // (En) sun", "lua // moon" (cf. "A lua está bonita! // The moon is beautiful!" and "Quantas luas tem Júpiter? // How many moons does Jupiter have?"), "Congresso // Congress";
- **NPROPs with apparently (and arguably) unitary extensions**, e.g. "(P) xadrez // (En) chess", "HIV", "gripe // flu".

Compounds. The treatment of groups of words as morphosyntactic units (resulting in compounds, marked by replacing spaces between their elements with the "=" symbol) is at one time imperative and dangerous. It is imperative because, otherwise, how could one tag e.g. "apesar/acerca/cerca" apart from preposition "de" as in "apesar/acerca/cerca de"? It is also dangerous because it is always difficult to establish clear criteria to decide whether to treat a given group as a compound. We chose the following ones:

- **non-analyzability**, which has already been implied, applying to "(P) apesar=de // (En) in=spite=of", "devido=a // due=to" and suchlike, and sanctions compounds (i) whose part-wise tagging is impossible or much too artificial, generating syntactically exceptional sequences of tags or (ii) whose (semantic) value seems not to be computable from the individual value of its elements;
- **trade-off**, recommending e.g. the consideration of many compound prepositions ("(P) antes=de // (En) prior=to", "depois=de // after", "perto=de // close=to", "longe=de // away=from", etc.) which could even be tagged as pairs of adverb plus preposition (introducing a complement of the corresponding adverb). However, we believe the latter possibility imposes an unnecessary cost on a subsequent syntactic analysis, since those are highly co-occurring items, expressing basic semantic relations (of time/space, among others) and generally behaving like any other one-word preposition;
- **non-productivity**, strongly correlating with non-analyzability and avoiding groups that, in fact, contain a currently productive syntactic-semantic structure, or rather, that are actually open-class. This criterion, for example, sanctions

“(P) a=cavalo // (En) on=horseback” and “*a=pé* // on=foot” while banning “de carro/ônibus/trem/etc. // by car/bus/train/etc.”

As one can see, our criteria are tenable, though a bit fuzzy, resulting in some of our highest inter-annotator inconsistency rates [5], in spite of some consistency-assurance devices we have devised (such as a central repository of compounds and candidates thereof). It is worth noticing that nearly as much as half the inconsistency is related to the creation of compound proper nouns, which is small wonder if one considers (i) how often proper nouns are in journalistic texts and (ii) how difficulty it is to determine how many proper nouns (only one or more) should be found in e.g. the following phrases: *“(P) Departamento de Computação do Instituto Tecnológico da Aeronáutica // (En) Department of Computation of the Airforce Technology Institute”*; *“Safári do Quênia // Kenia Safari”*; *“GP da Austrália de F1 // Australia’s Formula One Grand Prix”*; *“o SESC de São Carlos // São Carlos SESC”*.

Discontinuity. One important, perhaps novel feature of LW Tagset’s is the possibility of expressing discontinuity of morphosyntactic units, or rather, handling discontinuous occurrences of compounds, whether occasionally or necessarily so. That is realised by means of the complementary tags “[”, “…” and “]” (respectively denoting beginning, inner part and end of discontinuous unit) and seemed to be a good solution for two serious problems, namely:

- **“o mais ADJ/ADV possível”:** in Portuguese, structures like “(P) o(a) mais rápido(a) possível // (En) as soon as possible”, “o mais eficiente(s) possível // as efficient as possible”, “o mais à vontade possível // as at one’s ease as possible” are hardly susceptible, if at all, to analysis on a word-by-word basis (it is vital to notice that both “o” and “possível” are invariable, while inner adjectives are not). Even if we were to group “o mais” into a compound, how should we tag “possível” and it as independent entities? It seemed all the more appropriate to treat the whole “o=mais=possível” as a compound adverb and enable compound discontinuity. Hence the problematic structure can now be tagged thus: “o=mais_ADV|[**ADJ/ADV** possível_ADV|]”;
- **Compound Disruption:** perfectly eligible compounds have sometimes their usual continuity disrupted by extraneous elements inserted for emphasis or to prevent repetition of terms. Take e.g. the compounds “(P) apesar/antes=de_PREP // (En) in=spite=of/prior=to”. They may well happen to occur as “apesar/antes até mesmo de // even in spite of/prior to”, which can now be tagged thus: “apesar/antes PREP|[até=mesmo_PDEN de PREP|]”. One interesting example coming from our corpus is the following: *“(P) ...atingem níveis internacionais devido tanto à valorização interna quanto à valorização... // (En) ...reach international levels due not only to internal valorization but also to...”* tagged thus: “*...atingem níveis internacionais devido_PREP*|[*tanto_KC*|[*a_PREP*|]|+ *a_ART valorização interna quanto_KC*|] *a_PREP*|]|+ *a_ART valorização... // ...reach international levels due_PREP*|[*not=only_KC*|[*to_PREP*|] *internal valorization but=also_KC*|] *to_PREP*|]... ”

It is worth noticing that this device seems to be quite suitable to represent diverse binary coordinating structures (“(P) *tanto ... quanto/não só ... mas também // (En) not only ... but also*”, “*nem/já/ora ... nem/já/ora // either ... or/now ... now ...*”, among others).

4 Current and Future Work

We have developed MAC-Morpho, a 1.1-million-word Brazilian Portuguese reference corpus which shall be freely available on the Lacio-Web Project page (http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm). The total cost of tagging this huge corpus, including research on tagsets and tagging projects, corpus creation, writing the tagset manual, annotators' training, converting from Bick´s tagset to our tagset, weekly meetings with the annotators and revision took 11 months and involved 7 man month, 4 of them annotating the corpus. We ran two experiments to estimate inter-annotator agreement which presented kappa values in the .81–1.00 interval, namely 0.944 and 0.955, showing almost perfect agreement. The next steps will be a finer-grained correction phase of MAC-Morpho tackling the problems observed in the experiments and a tagset evaluation following [8].

References

1. Marques, N.C., Lopes, J.G.P.: A Neural Network Approach to Portuguese Part-of-Speech Tagging. Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado (1996) 1–9
2. Villavicencio, A., Viccari, R.M., Villavicencio, F.: Evaluating Part-of-Speech Taggers for the Portuguese Language. Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado (1996) 159–167
3. Aires, R.V.X., Aluísio, S.M., Kuhn, D.C.S., Andreeta, M.L.B., Oliveira Jr., O.N.: Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. Proceedings of SBIA'2000 (2000) 20–22
4. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus: Aarhus University Press (2000).
5. Aluísio, S. et al.: An account of the challenge of tagging a reference corpus of Brazilian Portuguese. Technical Report 188 – ICMC-USP (2003). Also Available at http://www.nilc.icmc.usp.br/~lacio_web/
6. Macleod, C., Ide, N., Grishman, R.: The American National Corpus: Standardized Resources for American English. Proceedings of the Second Language Resources and Evaluation Conference (LREC) (2000) 831–36
7. Galves, C., Britto, H.: A Construção do Corpus Anotado do Português Histórico Tycho Brahe: O sistema de anotação morfológica. Proceedings of PROPOR 99 (1999) 81–92.
8. Déjean, H.: How to Evaluate and Compare Tagsets? A Proposal. Proceedings of the Second Language Resources and Evaluation Conference (LREC) (2000). Also available at http://www.sfb441.uni-tuebingen.de/~dejean/