UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

An Account of the Challenge of Taggins a Reference Corpus of Brazilian Portuguese

Sandra Maria Aluísio
Jorge Marques Pelizzoni
Ana Raquel Marchi
Lucélia Helena de Oliveira
Regiana Manenti
Vanessa Marquiafável
Jorge Teles

Nº 188

RELATÓRIOS TÉCNICOS

São Carlos – SP
Fev./2003

Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

An account of the challenge of tagging a reference corpus of Brazilian Portuguese

Sandra Maria Aluísio
Jorge Marques Pelizzoni
Ana Raquel Marchi
Lucélia Helena de Oliveira
Regiana Manenti
Vanessa Marquiafável
Jorge Teles

NILC-TR-03-04
fevereiro, 2003

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Abstract

This article identifies and addresses the major issues faced in the manual morphosyntactic annotation of a huge corpus, named MAC-Morpho, a Brazilian Portuguese corpus of newspaper articles 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, make an account of how the annotation process was designed and conducted, including the results of the inter-annotator agreement evaluation for MAC-Morpho, 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 sizable linguistic resources, such as lexicons or grammars, and also to evaluate most computational analyzers. Usually, these treebank projects follow the Penn Treebank[1] approach, which distinguishes a POS tagging and a parsing phase each comprising an automatic annotation step followed by manual validation and correction. Recently, there have been several efforts to build gold standard annotated corpora for other languages than English such as French, German, Italian, Spanish, Slavic[2]. For Brazilian Portuguese, 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, of news from the "Editoria de Ciência e Tecnologia da Agência Brasil" [1,2], and (ii) the 104.966-word corpus[3] built from the corrected texts of the NILC corpus which is composed of 3 genres (newspaper, literature and textbooks) [3]. There are, although, several (Brazilian and European) Portuguese corpora **automatically** annotated with Bick´s [4] syntactic parser PALAVRAS[4], which is part of the AC/DC project[5] and has been constantly improved, with e.g. the addition of new proper nouns and compounds to the system's knowledge base. This parser is not freely available though, but Bick has gently applied it to several corpora used for scientific research. 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[6]. Lacio-Web, a two-year project launched in the beginning of 2002, tries to fill the gap with regard to base linguistic resources and tools for Brazilian Portuguese. It aims at compiling raw and annotated corpora and making them freely accessible for both **non-expert** people interested in Portuguese language and **expert** users who pursue theoretical and practical linguistic studies and develop computational linguistics tools and applications. In this report we present the rationale for building a 1.1 million-word corpus with manually validated morphosyntactic annotation, including the criteria for establishing the tagset, the automatic tool and filter used, how the annotation process was designed and conducted, results of the inter-annotator agreement evaluation, linguistic problems we faced in this work and directions for further work. The resulting annotated corpus (named MAC-Morpho) will be available in two versions: in annotators' format (one word per line followed by its tag) (see Apendix B) and in the XML-compliant format proposed by the Advisory Group on Languages Engineering Standards EAGLES [5][7].

---

[1] http://www.ldc.upenn.edu/Catalog/docs/treebank2/cl93.html
[2] http://treebank.linguist.jussieu.fr/
[3] http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.htm
[4] http://visl.hum.sdu.dk
[5] http://www.linguateca.pt
[6] http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm
[7] www.cs.vassar.edu/XCES

2. Methodology

2.1 Building the corpus

The 1,1 million words corpus was taken from a collection of texts from Folha de São Paulo (http://www.folha.uol.com.br/folha/) (1994) what give us high quality contemporary Brazilian Portuguese from different authors and domain. This corpus has been tagged by parser PALAVRAS and which also includes structural annotations (sentence, paragraph, title and subtitle). The criterion used to select data was random choice and to select only 10 from 22 issues was the opportunity to explore committee voting of taggers trained with texts from each of the 10 issues (see Figure 1). In order to choose 10 from 22 issues available in 1994 we used the following criteria: importance of the issues (given by their daily release and their number of tokens in one-year publication) what gave us the issues *Brasil* (3.877.787 tokens), *Cotidiano* (3.417.450 tokens), *Dinheiro* (3.254.763 tokens), *Ilustrada* (3.003.869 tokens); inclusion of issues presenting more elaborate syntactic structure – this gave us the issue *Mais*; grant a privilege to issues addressed to adult public – this excluded *Folhateen* and *Folhinha*; grant a privilege to certain domains, what resulted in the choice of *Agrofolha, Ciências, Informática, Esportes* and in the exclusion of *Imóveis, Veículos, Empreendimentos, Fovest, TVFolha*; varied proper nouns, what resulted in the choice of Mundo which brings a broad number of country and people names and historical landmarks; exclude issues whose vocabulary would be found at large in issues already chosen from the previous criteria (*Tudo* and *Revista* were excluded as their vocabulary can be found in *Cotidiano* and *Ilustrada*).

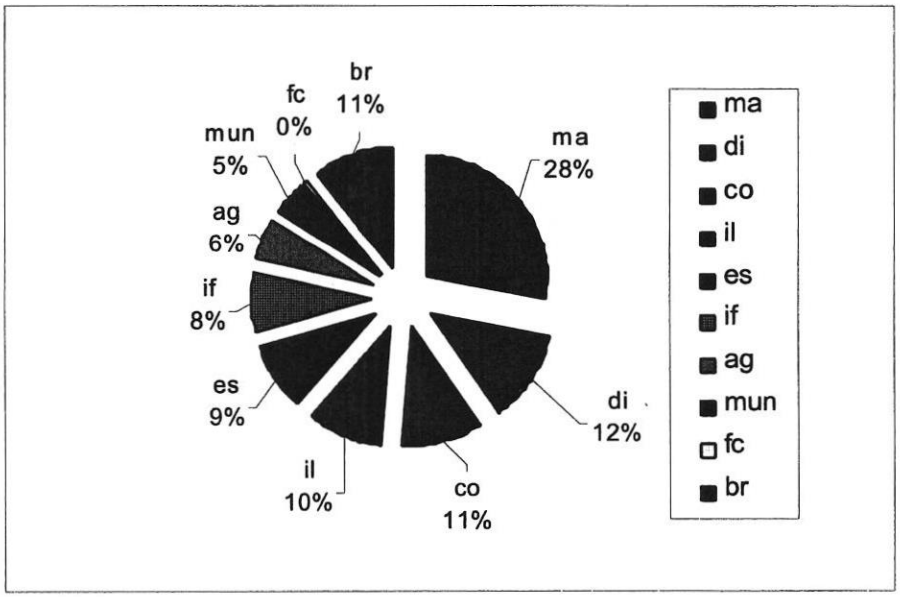

Fig. 1 The 10 issues chosen and their percentage in Mac-Morpho

2.2 Designing the tagset

The initial requirements for our tagset were simplicity and the ability to support subsequent syntactic parsing. Having that in view, we analyzed the Eagles recommendations for the Morphosyntactic Annotation of Corpora[8] and two of the more important tagsets designed for English (the Penn Treebank Tagset - with 36 POS tags -, and the two tagsets from the BNC project - the 61 basic tags from C5 and the 140 enriched ones from C7) and three other tagsets for Portuguese (NILC[9], PALAVRAS and Tycho Brahe [6] tagsets, respectively with 36, 14 and 48 tags).

Although there are already 2 tagsets for Portuguese (PALAVRAS and NILC), whose purpose is similar to that one we aim at, neither fulfills all the criteria we consider as essential to our project. These criteria will be discussed in the following subsection and have been employed by and large in Penn Treebank and Tycho Brahe projects. Even though the latter project also tackles Portuguese, it has been specifically designed to support a diachronic research and, perhaps due to this, end up with a conceptaully different tagset from ours.

2.2.1. Criteria, features and previous work

We identify below some of the leading criteria employed in the design of LW Tagset, exemplifying their application and presenting representative results thereof, usually in contrast with 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, ending up with 10 distinct pronoun tags. Were we to satisfy recoverability solely, 2 simple tags would do ("relative and non-relative pronoun"), to 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

---

[8] http://www.ilc.pi.cnr.it/

[9] http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm

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*[10] *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[11].

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

---

[10] 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. In the specific case of "falada", the impossibility is syntactical ("falar (de) // to gossip (about)" is not a transitive verb); but, in other cases, it may alternatively be semantic, e.g. when there is some addition to the meaning of the adjective that cannot be derived from the meaning of the verb, such as "assutado // timid" as opposed to "assustado // scared".

[11] Since "milhão/bilhão/quarto/etc." are only able to combine with a noun through a preposition (e.g.: "(P) um milhão de dólares // (En) one million dollars", "um quarto de pizza // one quarter of a pizza"), we understand they actually function as noun phrase nuclei.

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.

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.

2.2.3. Tagset Versions: a brief history

The Current Tagset

Since the beginning of its development, in July of 2002, LW Tagset has undergone cyclic revisions, being currently in its ninth version. At present it comprises 22 regular POS tags along with nine orthogonal complementary tags (see Appendix A). 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. Many of the complementary tags resulted from little puzzles found during the revision/correction process, e.g. related to (written) structures specific to journalistic texts, such as the insertion of dates, times, telephone numbers or sporting scores, and which are usually "expanded" into regular Portuguese when spoken. They also tackle actual linguistic phenomena, such as clitics and preposition-article contractions. Finally, there is the especially interesting case of discontinuity complementary tags, whereby one can denote that two or more non-adjacent tokens are to be analyzed as a single (morphosyntactic) unit just happening to be discontinuous. Discontinuity is rationalized and further discussed in Section 6.

Other Versions

In its first version, LW Tagset comprised 22 tags, namely 20 POS tags proper and the first two complementary tags, for contractions and clitics. It has been through much revision since, instances of which were the introduction and later (regretful) elimination of the complementary tag PASS (passive form), which could combine with both V (verb) and ADJ (in the case of a passive voice verb used adjectively, as in "(P) *o assunto discutido* // (En) *the subject discussed*"). As we will see in Section 6, the identification of the passive voice may be controversial, leading to mainly inter-annotator inconsistency. As a result, not only PASS was eliminated, but also the new indeterminate tag PCP was introduced.

In the following versions, besides detailing the tagset manual for better annotation as we tackled new challenges and consequently refined the very meaning of each tag, we introduced the POS tag CUR (currency) as well as the seven remaining complementary tags. At the moment, LW Tagset is stable.

3. Automatic annotation

Initially, the annotation of our 1,1 million-word corpus would be carried out manually by four annotators. However, after the Penn Treebank result that revision, whether followed by correction or not, is much more efficient than manual annotation[12], we also adopted this procedure, in which 4 annotators should revise and correct previously automatically tagged texts. This was only possible because we have a version of our corpus that had been gently tagged by the PALAVRAS parser, which generates a comprehensive morphological and syntactic characterization for each word of its input sentences. We verified that PALAVRAS's output is often more than sufficient to decide which LW tag is the most suitable in each case.

An additional advantage of PALAVRAS is that it does not generate explicit syntactic trees, but lists each word followed by its set of morphological and syntactic features encoding an (implicit) syntactic tree. This flattened, word-centered format allowed us to implement a very simple filter in Perl that processes each input line separately (a word and its features), according to 50 simple rewrite rules. As a result, we automatically obtained a new LW-tagged version of our corpus. Since there is some theoretical disagreement between these two annotation systems (especially related to compounds), this new version may present a bit higher error rate than that of PALAVRAS.

4. Hand-validation and Correction

As a preparation for the corpus revision and correction, the annotators were also engaged in the research of other existing tagsets for Portuguese and English. Thus, they were involved in all the LW Tagset definition process and aware of all the advantages and disadvantages of the other projects as well as the limitations of the LW Tagset itself. We intended that this involvement would benefit the annotation. Moreover, they went through a familiarization process with the environment (tagset manual, electronic dictionaries, central repository of compounds, problem base, text editor and tagged files) and practices (the annotation itself, maintenance of the repository of compounds and the problem base, and weekly meetings to discuss problems) to be used during the revision.

During revision/correction proper, the "original" automatic annotation is not modified, but "correction tags" are appended with the "#" symbol. Therefore, a corrected final version (see Appendix B) can be generated automatically, as well as keeping a record at hand for the annotators. This procedure facilitates determining how much has been corrected by the annotators, as well as pointing out errors in the automatic annotation (e.g. for tagger improvement) and evaluating the intra-annotator consistency.

Correction is not limited to changing tags, but also dealing with compound-related operations, such as breaking up miscompounds or grouping separately tagged tokens into recompounds. So we provide for line deletion (appending a "#" with no correction tag to the line to be deleted) and line rewriting (appending a sequence of "# <token>_<tag>" strings, each corresponding to a new inserted line replacing the original one). Figure 2a shows an instance of miscompound breaking, where "*pela=cúpula_ADV*" is broken up into "*por=**PREP**|+ a_ART cúpula_N*", and

[12] "Manual tagging took about twice as long as correcting, with about twice the inter-annotator disagreement rate and an error rate that was about 50% higher." (Marcus et al. 1993:pág)

Figure 2b exemplifies recompounding, where "*Valsa_N nº_N seis_NUM*" are grouped into "*Valsa=nº=seis_NPROP*".

```
e_KC
contribuições_N
de_PREP
campanha_N
pagos_ADJ #PCP
pela=cúpula_ADV    #por=PREP|+ #a_ART #cúpula_N
```

Fig 2 a. Example of miscompound breaking

```
" "
_
Falsa=Valsa=do=Anjo=Pornográfico_NPROP
" "
 _
',_'
adaptada_V      #PCP
de_PREP
" "
 _
Valsa_N         #Valsa=n°=seis_NPROP
n°_N            #
seis_NUM             #
" "
 _
```

Fig 2b. Example of grouping

5. Cost and Inter-annotator agreement

The whole cost of tagging this huge corpus, including research on tagsets and tagging projects, corpus creation, writing the tagset manual, annotators' training, filter development, weekly meetings with the annotators and the correction process itself took 11 months and involved 7 man month.

Apart from training, we ran two experiments in annotating parts of the corpus whose goal was to estimate (i) the average correction speed, (ii) the inter-annotator agreement and (iii) the percentage of annotation error classes. The first experiment involved all the four annotators in the task of correcting at most 100 sentences taken from the 10 newspapers sections composing our corpus. This experiment made use of tagset version 2, lasted 2 hours and took place just after the training phase, i.e. in second month of the correction period, which comprised 6 months. The second experiment took place 2 months later, also involving all the annotators in the task of revising at most 500 sentences taken from just 4 sections (the others had already been exhausted). This experiment used the last version of our tagset and lasted 4 hours.

In the second experiment, the average number of words corrected was 7.000, thus the average correction speed was 1.750 words per hour. This value is close to that of the Penn Treebank (2000 words per hour) which did not consider compounds.

In the 1995 AAAI Workshop on empirical methods in discourse, Isard and Carletta [7] proposed the Kappa Statistics as a measure of agreement for discourse analysis. This measure has been used by several researchers (Vander Linden and Di Eugenio [8]; Vieira [9], e.g.) as a test for a classification task in which several annotators assign items to one out of a set of classes. If there is total agreement among the annotators, Kappa will be 1 and if there is no agreement other than that expected to occur by chance, Kappa will be 0[13]. The Table 1 below presents a correlation between K values and inter-annotator reliability suggested by Rietveld and van Hout (1993)[14] apud Vander Linden and Di Eugenio [8].

Table 1. Correlation between K values and inter-annotator reliability

Kappa Value	Reliability Level
.00 – .20	Slight
.21 - .40	Fair
.41 - .60	Moderate
.61 - .80	Substantial
.81 – 1.00	Almost perfect

For the first experiment, the Kappa value was 0.944 showing almost perfect agreement; for the second, the Kappa value is even higher, 0.955. In all the calculations we ignored punctuation marks.

Figure 3 and 4 show that the higher disagreement percentage is caused by grouping and miscompound breaking (to be further discussed in Section 6).

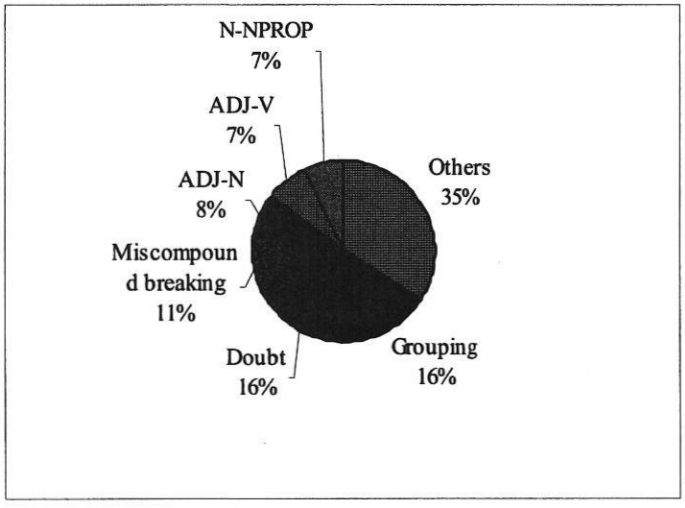

Fig. 3: Problematic tags in experiment 1

[13] Vieira (2002) presents a clear explanation on this measure.

[14] Rietveld, T. and van Hout, R. (1993). Statistical Techniques for the Study of Language and Language Behaviour. Mouton de Gruyter.

The cases of doubt are lower in the second experiment as expected after 4 months' revision. The usual confusion is also felt between N and ADJ (in spite of L-preservation, discussed in Section 2), in both experiments, and between V and VAUX (auxiliary verb), PREP (preposition) and ADV, and N and NPROP (proper noun), now in one experiment now in the other. It is worth explaining the 3% cases of hifen tag in the second experiment: this occurred due to formatting errors in the original file.

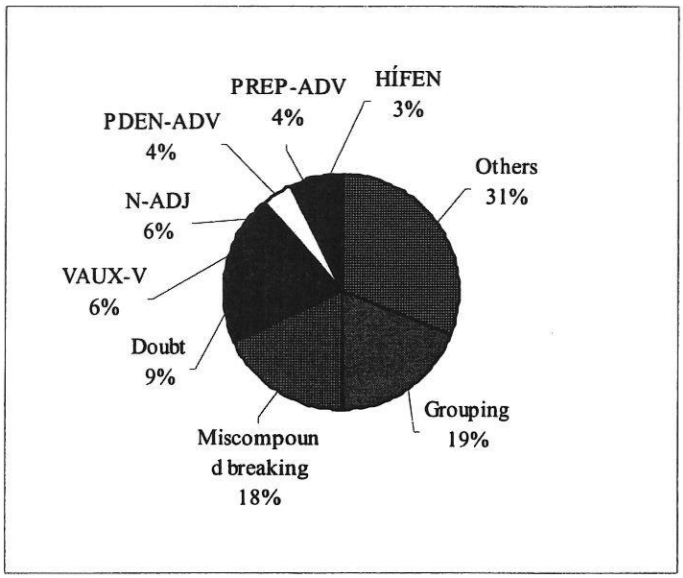

Fig. 4: Problematic tags in experiment 2

6. Some emblematic linguistic challenges

In this section we will analyze some interesting cases amongst the linguistic problems we faced in this work.

Auxiliary Verb:

Distinguishing between auxiliary and main verbs seemed desirable since it would make it easier to identify predicate nuclei and clause boundaries. However, except for a few too consensual cases (perfect-aspect auxiliaries "(P) *ter/haver //* (En) *to have*", passive-voice auxiliary "*ser // to be*", continuos-aspect auxiliary "*estar // to be*", etc.), there is plenty of room for doubt in this task. What is worse: as usual, all the grammars we consulted presented no definitive criteria but lists of auxiliary verbs without much justification.

In order to ensure consistency in auxiliary verb annotation, we developed the following criterion (which may be disputable but covers all consensual cases and most that are usually regarded as auxiliary verbs): a verb is considered an auxiliary if and only if it does not prevent its supposed main verb undergoing the usual commutation of voice with (relative) preservation of meaning.

Passive Voice:

The option to identify auxiliary verbs naturally entails also identifying the passive voice. Initially, we intended to extend voice analysis to noun modifiers, as in "(P) *as meninas nascidas_ADJ em 1980* // (En) *girls born_ADJ in 1980*", where there is an adjective proper, versus *"os quadros produzidos_ADJ\PASS em 1500 // the pictures produced_ADJ\PASS in 1500"*), where there is a passive-voice verb in adjectival position.

However, we were unable to find or develop any criteria leading to consistency in cases such as:

"(P) *O brinquedo está quebrado* // (En) *The toy is broken."*
"O trabalho é/está baseado em dados reais // The work is based on real data."
"A menina é/está assustada // The girl is scared/timid."

KC *vs.* ADV:

Many elements that are traditionally treated as coordinating conjunctions (KC) have adverbial behaviour, perhaps to the extent of making its traditional status disputable. Among such elements are some conjunctions of contrast ("(P) *no=entanto/entretanto/etc.* // (En) *however/nevertheless/etc.*") and logical consequence ("*portanto/por=conseguinte/etc.* // therefore/consequently/etc."), which not only present some mobility inside the conjunct but also may co-occur with other coordinating conjunctions (as in "(P) *Ofendeu nosso irmão e, portanto, a família toda* // (En) *He/She offended our brother and thus our whole family"*).

In order to generate well-behaved syntactic structures, we choose tagging such elements as adverbs if and only if they do not function as main connecting devices, i.e, when in combination with other coordinating conjunctions. Thus: "(P) *Ofendeu nosso irmão e_**KC**, portanto_**ADV**, a família toda* // (En) *He/She offended our brother and_**KC** thus_**ADV** our whole family"*

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*. More technically speaking, as the extension of the proper noun is unitary, it is not possible to state its intension, namely the set of all discernible properties common to all members of the extension of a word.

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*", "*Instituto de Ciências Matemáticas e de Computação de São Carlos // Institute of Mathematic Sciences and Computation of São Carlos*");
- **metonymical NPROPs**, e.g. "(PBr) *gillette* // (En) *(brand of) razor blade*", "*band-aid*", "*danone // (brand of) yoghourt*", "*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*".

It is worth mentioning that cases (iii) and (iv) are especially trying in that, even after some time's discussion of specific instances, annotators will not reach agreement more often than in other cases.

Compounds

The treatment of groups of words as morphossyntatic 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 choose 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/avião/etc.* // *by car/bus/train/plane/etc.*"

As one can see, our criteria are tenable, though a bit fuzzy, resulting in some of our highest inter-annotator inconsistency rates (Figure 4, under "grouping" and "miscompound breaking"), 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 de Ciências Matemáticas e de Computação de São Carlos // (En) Department of Computation of the Institute of Computation and Mathematics of São Carlos)[15] *";*
"Safári do Quênia // Kenia Safari";
"GP da Austrália de F1 // Australia's Formula One Grand Prix";
"o SESC de São Carlos // São Carlos SESC".

The correct resolution of these cases involves, among other things, real-world knowledge that much too often is not available to annotators. Case (d) is still more interesting because the very status of "*SESC*" as a proper noun, usually indisputable, is not so much so there.

Next we will analyze illustrative instances of the application of the above criteria to some problematic groups.

"quem quer que":

The common phrases "(P) *quem/onde/o que/quando/etc. quer que* // (En) *who/where/what/whenever etc.*" are all-worthy compounds, being tagged as either ***PRO_KS*** or ***ADV_KS*** according to the nature of the first element of each expression (either pronouns or adverbs). The crucial criteria, in this case, are (i) non-analyzability (tagging "*quer // to want*" as a verb would lead to absurd syntactic and semantic interpretations — also involving the consideration of an elliptical agent in some cases — or simply non-grammatical readings) and (ii) non-productivity (all the elements combining with "quer que" are closed-class and, moreover, invariable).

"seja quem for (que)":

Especially after deciding for the status of "(P) *quem quer que/etc.* // (En) *whoever/etc.*" as compounds, it may seem natural to treat "*seja quem for (que) // whoever it is (that)*" alike, and even tempting, as that is incidentally a structure of no trivial analysis. However, productivity is never to be overlooked: phrases such as *"esteja onde estiver // wherever you are", "faça o que fizer // whatever you do", "tome qualquer precaução que for/tomar // whatever precaution you take"* — which are even variable, e.g. *"estejamos onde estivermos // wherever we are"* — evince that there is a productive structure there, abstract though it may be. And, in principle, all productive structures must be accountable for by a language model (if not, that will be purely out of deficiency).

That is exactly why non-productivity is "the" criterion to stick to, with priority over all others, and why we simply had to find an explanation (= analytical tagging + hypothetical parsing) for the structure at issue. Our conclusions are the following:

> in these structures, "*quem*", "*onde*", "*o=que*", etc. must be regularly tagged PRO-KS(-REL) or ADV-KS(-REL);
>
> because these devices introduce common subordinate clauses (*"quem for (que...)", "onde estiver", "o que fizer", "que tomar"*), to be normally analyzed; and

[15] English translations will not do justice to the difficulty in the corresponding Portuguese examples, since English have additional ways of building noun phrases to Portuguese's almost ubiquitous use of prepositions.

what is odd about the structure is that the subordination of the referred subclauses (*"[seja [quem for]$_{COPULA\_COMPL}$]$_{ADV}$"*, *"[esteja [onde estiver]$_{ADV}$]$_{ADV}$"*, *"[faça [o que fizer]$_{DIR\text{-}OBJ}$]$_{ADV}$"*, *etc.*) to yet another clause is realized solely by the subjunctive mood of verbs (*"seja"*, *"esteja"*, *"faça"*, *etc.*) and inversion.

"por mais que":

Another apparently eligible compound is "(P) *por mais que* // (En) *much as (=although)*". It is, however, an analogous case to "*seja quem for*". One again, non-productivity is not satisfied, it being enough to consider e.g. "*por mais comida que coma // much food as he may eat*", "*por poucas palavras que diga // few words as he may say*", "*por menos contente que esteja // little happy as he may be*", "*por mais cedo que se chegue // much early as one may arrive*", "*por muito que lute // much as one may fight*" e "*por muita gente que venha*".

We accordingly decided to consider the pair "*por ... que*" in this structure as a discontinuous (see below) concessive subordinating conjunction, being thus tagged "*por_**KS**\[... que_**KS**\]*". Those elements surrounded by this conjunction are regularly tagged, which means that "*mais/menos/muito/pouco/etc.*" receive ADV, PROADJ or PROSUB, depending on the case.

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[16] 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 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... //*

[16] Complementary tag "..." is but a theoretical possibility.

(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).

The numeral phrase and compounds "cerca=de", "menos=de" and "mais=de":

During the annotation process, we had to tackle the syntactic analysis of such phrases as "(P) *mais/menos/cerca de dez pessoas // (En) more/less than/about ten people*". The first issue therein was to determine the nuclei of those phrases. It seemed very artificial to consider "*mais/menos // more/less*" as nuclei, instead of "*pessoas // people*". Moreover, even that could not be a definitive solution, since it was beyond cogitation to treat "cerca // about" likewise.

In search of a solution, we looked into related phrases like "(P) *de cinco a dez pessoas // (En) from five to ten people*", "*entre cinco e dez pessoas // from five to ten people*" and "*acima de cinco pessoas // above five people*". Then we noticed that:

I. each of those phrases is a noun phrase beginning with a preposition (!) ...
II. ... internal though to the phrase and thus somehow subordinated to its nucleus, ...
III. ... which is inevitably "pessoas".

Therefore, we postulate the *numeral phrase* syntactic category, possibly embedded in noun phrases and with syntax of its own. This supports our hypothesis that "*mais/menos=de // more/less=than*" and "*cerca=de // about*" are compound prepositions, in perfectly parallel structures to those created with "*acima=de // above*".

7. Current and Future Work

We have developed MAC-Morpho, a 1,1 million-word Brazilian Portuguese reference corpus which shall be freely available from the Lacio-Web Project page[17]. 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 proper 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, respectively 0.944 and 0.955, showing almost

[17] http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

perfect agreement. The next step will be a finer grained correction phase focusing on the problems occurred in the experiments.

Acknowledgements

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

References

1. Marques, N. C., Lopes, J. G. P.: A Neural Network Approach to Portuguese Part-of-Speech Tagging. In: the *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. In: the *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. In: 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. 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.
6. Galves, C., Britto, H.: A Construção do Corpus Anotado do Português Histórico Tycho Brahe: O sistema de anotação morfológica. In: Proceedings of PROPOR 99, (1999) 81-92.
7. Isard, A., Carletta, J.: Replicability of Transaction and Action Coding in the Map Task Corpus. In AAAI Spring Symposium on Empirical Methods in Discourse Generation and Interpretation, (1995) 60-66.
8. Vander Linden, K., Di Eugenio, B.: Learning Micro-Planning Rules for Preventative Expressions, In Proceedings of the Eighth International Workshop on Natural Language Generation, (1996) 11-20.
9. Vieira, R.: How to evaluate systems against human judgment on the presence of disagreement? In: *Anais do Encontro Preparatório de Avaliação Conjuntado Processamento Computacional do Português*, (2002). Available in: http://acdc.linguateca.pt/aval_conjunta/Faro2002/Renata_Vieira.pdf

Appendix A: Final Tagset

TAG	DEFINITION	CLOSEST TRADITIONAL POS
ADJ	open-class noun modifier	adjective
ADV-KS-REL	relative subordinating adverb	relative pronoun
ADV-KS	non-relative subordinating adverb	interrogative adverbs in noun clauses
ADV	non-subordinating adverb	adverb
ART	article	(ditto)
KC	coordinating conjunction	(ditto)
KS	coordinating conjunction	(ditto)
IN	exclamation	(ditto)
N	open-class noun phrase nucleus	noun
NPROP	proper noun	(ditto)
NUM	numeral as a noun modifier	cardinal numerals and inflections of "meio"
PCP	past participle or adjective zero-derived therefrom	past participle
PDEN	emphasis/focus	emphasis/focus
PREP	preposition	(ditto)
PROPESS	personal pronoun	personal pronoun
PRO-KS-REL	relative subordinating pronoun	relative pronoun
PRO-KS	non-relative subordinating pronoun	interrogative pronoun in noun clauses
PROSUB	non-subordinating pronoun as a noun phrase nucleus	pronoun
PROADJ	non-subordinating pronoun as a modifier	pronoun
VAUX	auxiliary verb	(ditto)
V	non-auxiliary verb	(ditto)
CUR	currency symbol	NA

Complementary Tags	Meaning
|EST	foreign
|AP	apposition
|+	contraction/enclitic
|!	mesoclitic
| [	beginning of discontinuous unit
|...	inner part of discontinuous unit
|]	end of discontinuous unit
|TEL |DAT |HOR	phone no. date time

|DAD	formatted data not falling into the above categories

Appendix B: Corpus Sample in the Lite Format (annotator´s format)

```
Animais_N
gerados_ADJ  #PCP
por_PREP     #PREP|+
o_ART
cruzamento_N
de_PREP
limousin_ADJ#N
com_PREP
nelore_N
serão_VAUX
vendidos_V   #PCP
em_PREP
Botucatu_NPROP
(_(
SP_NPROP
)_)
._.

Temporada_N
tem_V
duas_NUM
provas_N
em_PREP
abril_N
e_KC
maio_N
,_,
uma_PROSUB
de_PREP              #PREP|+
elas_PROPESS
escolhe_V
equipe_N
para_PREP
o_ART
Mundial_ADJ #NPROP
._.
```