Evaluation of parsed corpora: Experiments in user-transparent and user-visible evaluation

Diana Santos*, Caroline Gasperin†

* SINTEF Tele og Data
Pb 124, Blindern, NO-0314 Oslo, Norway
Diana.Santos@sintef.no

†Faculdade de Informática, PPGCC, PUCRS
Av. Ipiranga, 6681, Prédio 16, 90619-900 Porto Alegre, Brazil
caroline@inf.pucrs.br

Abstract

In the present paper, we describe and discuss the evaluation of parsed corpora, namely the ones that are available on the Web for querying in the AC/DC project. The paper has two parts: the first one suggests a set of different evaluation parameters and measures that are much more illuminating than commonly used simple precision measures, while the second evaluates the parsed corpus for a particular task -- that of automatic thesaurus building. The two evaluations are thus complementary, in that, in Gaizauskas (1998) terminology, the first is a typical user-transparent evaluation, while the second is user-visible.

1. Introduction

There is at present a large activity as far as parser evaluation is concerned, witnessed among other things by the workshop "Towards improved evaluation measures for parsing systems" at the present conference.

We are concerned here with the closely related subject of *parsed corpora evaluation*, which brings, however, a different perspective into the picture. In fact, although a parsed corpus can be seen as a frozen picture of a parsing system, it has usually a life of its own, and a set of users, and uses, which are different from those of the parser itself. In addition, many of the parsed corpora presented as such, or as treebanks, include human revision and therefore problems and capabilities beyond those provided by a parser itself.

Santos and Bick (2000) presented the AC/DC project[1], a Web service giving access to Portuguese parsed corpora using the PALAVRAS parser (Bick, 2000), and mentioned the need to make user studies to evaluate its usefulness and the quality of the underlying annotation.

We believe there is too little work on the evaluation of language resources in themselves (as compared to programs, systems or tools), although it might be argued that the first kind should be easier to evaluate than the second. Santos and Rocha (2001) attempted to evaluate a large corpus as far as structure and tokenization was concerned. Here, we try to go a step further and look at syntactically annotated corpora.

2. Goal and outline of the paper

The primary motivation for the paper is the need to provide users of the AC/DC service with rigorous information of what is being supplied, and what the shortcomings are that are (vaguely) known to exist in the material.

Gaizauskas (1998) has suggested to bring the user into the evaluation of NLP applications. He distinguishes between *user-transparent* evaluations, that look in terms of input and output of a particular computational-linguistic task, which may not make sense for a external user, and *user-visible* evaluations where one is measuring success relative to a particular task a user understands and is involved with.

In the context of corpora resources, the typical user-transparent question concerns the quality of the actual tagging and parsing, while user-visible evaluation depends mainly on what a user is supposed to do with the parsed corpus, and how directly its quality matters for that task.

In this paper, we suggest a series of criteria for the first kind of evaluation and measure some of them; for the second kind, we investigate the task of automated thesaurus building (Grefenstette, 1994) following Gasperin's (2001) work for Portuguese.

We are most grateful to Eckhard Bick to have supplied his PALAVRAS, and would like to emphazise early on that what we are presenting here is *not* a parser evaluation. In fact, our parsed corpora have been created using several different versions of the parser (none reflecting its today's performance) and, besides, the final rendering of the parsed corpora amounted to differences in around 20% of the tokens, as detailedly explained in Santos and Bick (2000), which means that the AC/DC project in itself "added" many parsing options, and possibly many mistakes as well.

Still, the parsed corpora exist and are being actively used by an increasing user community. Therefore, they deserve to be evaluated in their own right and qualified so that they can be improved and its improvement measured, something which so far has not been possible to do in a systematic way.

We provide here a short description of the corpora used for the present paper:

ENPCANOT (v.2.3) is a corpus of translations of English fiction texts into Portuguese, a subset of the Portuguese part of the ENPC corpus (Johansson et al., 1999; Santos and Oksefjell, 1999) containing around 70,000 words. It was manually revised by the ENPC

[1] See http://cgi.portugues.mct.pt/acesso/.

team and contains texts in the European and Brazilian variants of Portuguese.

EBRANOT (v.3.4) is a part of the Borba-Ramsey corpus, distributed by the ECI/MCI iniciative, and contains exclusively Brazilian text in several genres: literary, newspaper, scientific articles and law, amounting to 700,000 words.

NATPANOT (v.2.6) is a corpus of 8 million words of newspaper text (1991-1994) in European Portuguese.

FOLHANOT is the first million of a newspaper text corpus in Brazilian Portuguese, currently in development by the AC/DC project. It is a proper subset of the SCANOT corpus, compiled by NILC.

3. Annotation quality

Although one could in principle be interested in all aspects having to do with an annotated corpus, such as: Is there sufficiently encompassing documentation? Is there a formal definition, in the form of e.g. a DTD? Does the corpus conform to it? Has the corpus been validated by a third party? Has it been evaluated? etc. etc., we will be here solely concerned with what is central to the parsing issue.

3.1. What should a parser do?

By informing others that a corpus is parsed, we implicitly state at least the four tenets: 1) The text units (tokens) have been recognized and assigned to their right category (lemmatization and PoS tagging); 2) MWE have been identified (tokenization); 3) Morphological information has been made explicit (morphological analysis); and 4) Syntactic constituents and relations have been identified (couched, depending on the theoretical inclinations, as constituency, functional and/or dependency structure). Additionally, other kinds of information can also be present in parsed corpora, such as named entity classification, anaphoric dependencies or rethorical structure, which we will disregard here since they are absent from the AC/DC corpora.

Not all these tasks are equally relevant or well defined, and not all the problems that are to be solved equally frequent. In addition, there are strong dependencies between these tasks, with the vexing property that each requires a different unit of measure. We will try to describe these problems in detail with the help of the AC/DC corpora. But first we turn to the problem of assessing separately the different kinds of information.

3.1.1. In which level is one particular phenomenon handled?

In many cases it is up to the parser developer in which linguistic level – better, in which way – a particular distinction made in language should be encoded in the output of the parser. This should refrain one to evaluate levels independently, especially when comparing different parsing approaches.

One examples is the choice between encoding a particular syntactic difference as PoS or as constituent function. In *Três quartos do hotel foram ocupados pela polícia* (three quarters/rooms of the hotel were taken by the police) one can represent the difference by assigning the PoS noun to *quartos* in one interpretation and the PoS numeral in the other. Alternatively, one can have both parses tagging quartos as noun at the PoS level, but individuated by their function inside the NP *Três quartos do hotel* – having either *quartos* (rooms) or *hotel* as NP head.

Another encoding alternative is between PoS or constituent type: In *Os pobres saíram* (the poor left), *pobres* may be assigned the PoS noun and assigned head of the NP, or the PoS adjective and still head of the NP, both conveying the same thing (though with different underlying theories).

The same liberty at making distinctions can be seen in the three sentences *Ele está de volta* (He is back), *De volta da mãe, ele apressava-se* (Around mother, he hurried) or *Comprou o bilhete de volta* (He bought the return ticket), where a parser can give the same PoS, viz. preposition noun, to the three instances of *de volta*, but separate them by function (e.g. by AJP, AVP and PP), or actually perform three different tokenizations as well: "de volta", "de volta de", and "de" "volta".

Examples could be multiplied at will – what is relevant is the need to understand the parsing scheme in order to distinguish wrong parses from systematic ways of dealing with a particular phenomenon.

3.1.2. Categorial ambiguity

The first requirement or expectation when facing a parsed corpus is that words that are categorially ambiguous out of context are assigned their right part of speech. But measures such as percentage of right PoS assignment have long been shown to be inappropriate (Santos, 1999), because they do not take into account the difficulty of the problem, both from a macro and from a microperspective:

In fact, ca. 90% of the words in a text (66% of the types) are unambiguous (for example, most of those that belong to a closed set such as prepositions, conjunctions, personal pronouns, negative adverbs, etc.)[2]. In addition, if for all wordforms that belong much more frequently to one PoS than the other the more frequent label is assigned, overall one gets more than 95% of PoS labels right. However, this is no measure of the quality of PoS tagging, giving that, if such a procedure were followed, gross syntactic incorrections might occur, such as the sequence of two syntactically incompatible tags...

One should compute, for each potentially ambiguous form present in the corpus, what the difficulty and the information-theoretic gain is of deciding what is their PoS, in order to be able to measure the job done by the parsing procedure.[3] For each pair of <wordform, PoS> could then precision and

[2] These numbers are based on old studies regarding *major* PoS, done for Portuguese (Medeiros et al., 1993; Santos, 1996). In Table 5 ahead, concerning a hundred PoS distinctions, the number of unambiguous forms is only slightly above 50%.

[3] It is true that one should also consider the (few) cases where the only PoS assigned is wrong (and which may come from guessing about unknown words, or even from wrong dictionary entries). However, this should not, in our view, be brought to the same count as all the unambiguous words whose PoS was right by simple dictionary lookup.

recall be measured (see Hindle and Rooth (1993) for the need to have different PR-measures for each choice).

Table 1 presents some of these figures in the small ENPCANOT corpus, for wordforms ambiguous between verb and noun readings. The data column presents correct noun readings – wrong noun readings – wrong verb readings – correct verb readings. The PR column presents noun precision, noun recall, verb precision and verb recall.

wordform	Data	PR
espera	10 0 0 2	1.0 1.0 1.0 1.0
casa	95 0 1 1	1.0 .989 1.0 0.5
ser	7 2 0 147	.77 1.0 1.0 .986
volta	36 0 0 1	1.0 1.0 1.0 1.0
sentido	7 0 4 4	1.0 .636 .5 1.0
ouvido	1 0 0 6	1.0 1.0 1.0 1.0
jantar	10 1 2 5	.909 .833 .714 .833
comida	10 0 1 1	1.0 1.0 .5 .5
gosto	7 0 0 6	1.0 1.0 1.0 1.0
vinda	1 1 0 2	.5 .5 1.0 1.0

Table 1: Evaluating noun/verb disambiguation

It is at once obvious that no averaging of these numbers will do, since for different forms (or contexts) the parser will do better for "nounness" or "verbness". The table above just shows that we have inspected 194 rightly analysed nouns, four verbs incorrected labelled as nouns, eight nouns incorrectly classified as verbs, and 175 correctly identified verbs. Noun precision (.979=194/198) will not be a function of the individual noun precisions, nor the other values will: verb precision .956, noun recall .960, verb recall .977.

Things get even more difficult when realizing that there are more complex disambiguating tasks also measurable in noun or verb precision, namely ambiguity with other parts of speech. Table 2 illustrates similar calculations for other PoS pairs or trios (only precision/recall regarding the two first PoS is presented, though taking into consideration all analyses).

wordform	Data	PR
desses*(gram/V)	8 2 0 2	.8 1.0 1.0 .5
sobre(gram/V/N)	105 0 3 3	1.0 .97 .5 1.0
suas(gram/V)	84 0 2 2	1.0 .98 .5 1.0
alto(ADJ/ADV/N)	13 1 3 6	.93 .93 .67 .86
claro(ADJ/ADV)	24 21 1 4	.53 .96 .8 .16
quartos(N/ADJ)	4 0 0 1	1.0 0.0 0.0 1.0
creme(N/ADJ)	3 0 0 2	1.0 0.0 0.0 1.0
presentes*(N/ADJ)	3 2 0 9	0.6 1.0 1.0 .82
tarde(V/ADV/N)	1 1 1 41	0.5 0.5 .98 .98
fora(V/ADV)	54 3 7 36	.95 .89 .84 .92

Table 2: Evaluating other PoS disambiguation tasks

It should in any case be noted that even a seemingly simple task as deciding for PoS is marred by the difficulty, alluded before, of identifying the correct level where information is encoded – and the converse, which level to assign an error if it is *not* conveyed. For example, consider the following phrases:

- *à espera* (waiting) constitutes an adverbial phrase (though the word *espera* is related to the noun *espera*, waiting)
- *ao largo* (at a distance) also works as an adverb (the word *largo* is not related to the noun *largo*, square), and
- *a seguir* (next) is only metaphorically related to the verb *seguir* (follow), being used in much wider contexts than an infinitive phrase, namely as a complex adjective or adverb

So, should one consider the simple assignment of respectively noun, noun and verb to *espera*, *largo* and *seguir* a right PoS assignment? If the distinction were encoded in other parts of the analysis, maybe yes. If not – where to measure it? The easiest way would be to remove these cases simply from PoS accounting, and expect them to be rewarded (or punished) at the right level. But let us note the lack of a golden rule for measuring and encoding these matters: there is no universal or near universal consensus on what the text units should be (words or multiword expressions). So, if one does not want to incur in a judgement of the underlying grammatical theory -- then we would be actually comparing two different parsing approaches and not a parsed corpus in itself -- there are only two ways left. The first is using the extensional limits provided by the parser and consider the output right when no other alternative is possible. The second is to use, besides right and wrong, a third category in our precision and recall computations, to mean that the relevant distinction is or should be encoded at a different level (and then reward it at that level).

One of the places where this is more obviously reflected is in tokenization. See Santos and Bick (2000) for an illustration of the amount and kind of differences.

3.1.3. Lemmatization and morphological analysis

In many cases, lemmatization is trivial after PoS assignment,[4] and therefore should not get more credit for the parser, but not always, because of intracategorial ambiguity. This is especially common for verb forms in Portuguese, but also possible in nouns, as illustrated in Table 3:

wordform	possible lemmas
fora(V)	ser ir
vendo(V)	vender ver vendar
vira(V)	ver virar
revista(V)	rever revistar
costas(N)	costa costas
graças(N)	graça graças
vimos(V)	ver vir
amara(V)	amar amarar
assente(V)	assentir assentar

[4] Assuming that the underlying morphological analyser is reliable, which is obviously a simplification, especially in the case of unknown words. For a proportion of these in Portuguese text see Reis (1993); for a study of the performance of PALAVRAS in this respect see Bick (1998).

lido(V)	ler lidar

Table 3: Lemma ambiguity

The disambiguation of morphological features, when they are not defined by lemma and PoS, is yet another task on which to measure the performance of a parser (and/or the quality of annotation of a parsed corpus). Clear examples are the pervasive ambiguity (for all but the most irregular verbs) between

- future of subjunctive and infinitive forms;
- first and third person of imperfeito (in both indicative and subjunctive moods);
- perfeito and pluperfect tense in the third person plural;
- perfeito and present in the first person plural[5]

This also holds for gender of nouns such as *capital*, *moral*, *presidente* and those ending in *ista*, as well as for gender of a large class of invariant adjectives. A less considerable task is number disambiguation for a few nouns and adjectives. Finally, one further non trivial task of a parser is to assign gender (and number) to a proper noun (something not necessarily obvious even for a human being, see Afonso et al. (2002a) for discussion).

It is arguable whether gender and number of (non-lexically determined) pronouns should be considered as a morphological disambiguation task. We will not consider it here, although all pronoun instances are marked in the corpora as M/F (both genders possible).

In fact, one important fact regarding morphological ambiguity in the present parsed corpora is that most of it is simply not resolved, which means that a large number of wordforms still carry portmanteau labels (15%, 11% or 10% of all the forms not classified as invariant). Just to give a more precise idea of what this means in practice, let us look at the disambiguation of presente and perfeito in the first person plural in the EBRANOT corpus. In the 1745 cases marked present and/or perfeito in that person, the distribution is as displayed in Table 4.

Tense	Total	Ambiguous
presente	806	52
perfeito	168	44
presente/perfeito	771	771
Total	1,745	867

Table 4: Disambiguation of tense form

While almost 60% of the forms have only one label assigned, a quick inspection of both presente and perfeito cases shows that the vast majority of them was already unambiguous from the start (belonging to those verbs having distinct forms). So, in practice, the disambiguation task was only done in 96 forms out of 867 (11% of the cases).

Following the same procedure to analyse this kind of disambiguation task as used for PoS: Of 44 forms analysed as perfeitos, 4 are wrong (should be presente) and 40 right, thus yielding a precision of perfeito recognition of .91. As for the 54 forms classified as presente, 13 are wrong (3 of them featuring as well a wrong lemma, one of which due to a spelling error, so only 10 are actually perfeitos), 35 are right, and 6 are possible in the two interpretations (even consulting the largest possible context). In order to simplify the present computations we stipulate, in this case, three wrong and three right (three perfeitos and three presentes). Thus we get .70 precision in identifying presente, and .90 coverage, while we have .75 coverage for perfeito identification.

3.1.4. Syntax proper

To talk of a parsed corpus instead of a tagged one, larger elements than words (or basic units) have to be identified, and (some of) their functions have to be revealed. This is the more complex part of the parsing work, and it is also the one which requires more complicated assessment procedures, even if one is simply evaluating *one* parsed corpus and not competing schemes of annotation (as concerns Black et al. (1991), Lin (1995) or Carroll et al. (1998)).

As far as we know, there is no fixed number of syntactic distinctions that one can use as a measure, and syntax, as opposed to morphology, is still a sparsely exploited area. There is no other way, it seems, at least for the time being, than to conform to the theory obeyed by the parser and, inside its limits, test what is right and wrong. One has to list the possible analyses contemplated (forget those that were not) and, in light of the alternatives, decide whether the result is the best possible. For the AC/DC corpora, the underlying theory is dependency based, so there is no direct way to define constituents, and there is quite a large number of cases where attachment is left unspecified. Besides, and as was the case for morphological information, there is a considerable number of alternative function tags that have not been disambiguated.

But still, for each verb which admitted of an object one could compute precision and recall of object detection; for each verb which admitted an object, one could compute the PR figures; for each ditransitive verb of the form NP PP one could check them, and so on. Conversely, for all sentences one could check appropriate main verb detection, as well as (apparent) right argument structure.

Again, one has to be careful about what is the domain of possible/wrong categories, even when function labels are assigned to every word. (All syntactic information available in the AC/DC parsed corpora is through function labels.)

[5] Due to different spelling conventions, this applies only for verbs ending in *er* or *ir* for European Portuguese, but for almost all verbs in Brazilian Portuguese. As far as we know, this was not taken in consideration in the automatic analysis, resulting in a much larger number of initially ambiguous forms -- and actually incorrect portmanteau tags -- in the European Portuguese texts.

Corpus	Size	PoS ambiguity	Lemma ambiguity	Morphological ambiguity
ENPCANOT	72,431 12,886	29,531 730	3,264 41	13,063 905
EBRANOT	722,715 60,118	348,576 4,419	39,974 123	164,760 6,654
NATPANOT	6,295,653 167,206	3,223,063 11,534	448,916 420	1,570, 102 18,395

Table 5: Some extensional measures of disambiguation need for three different corpora

In fact, PoS may uniquely determine function, as is the case with articles, always assigned the function label N>. Also simple PoS sequences such as preposition (article adjective*) noun, result in the noun necessarily getting the label P< and all intervening articles and adjectives the label >N.[6] This, incidentally, constitutes respectively 49%, 48% and 52% of all words classified as nouns in the corpora we are dealing with.

3.2. Relevant characteristics of a parsed corpus

After detailing the problems and before suggesting measures, we would like to note that a corpus, no matter how large, has a fixed vocabulary, so that quality features for each word can be exhaustively computed, as well as the difficulty involved in parsing it (prior to parsing).

So, for each wordform occurring in the corpus one can know its span (the set of different possible analysis).[7] It is therefore possible to give a first measure of the parsing difficulty of a corpus by presenting statistics like the percentage of ambiguous wordforms. It is important to stress, if one is *comparing* (and not only evaluating) corpora, to realize that different corpora may offer different challenges to syntactic analysis.

We can use, for this estimation, both internal and external criteria. Internal criteria are what the corpus in itself reveals, having the number of forms assigned different analyses as one measure of the disambiguation difficulty present in the corpus. This is, obviously, a measure by default: All ambiguous forms that have been disambiguated and have been found to occur in only one way are counted as unambiguous... but note that possible error is neither computed as well (forms with one analysis in the corpus pair with unambiguous forms).

External criteria would use other sources of probing, like morphological analysers and lemmatizers. Ideally, the ones used by the parser itself.

Table 5 gives, for three different corpora, the following figures, obtained by internal criteria: sheer size in words, number of categorially ambiguous word forms, number of intracategorially ambiguous word forms as far as different lemmas are concerned, and number of intracategorially ambiguous word forms as far as morphology is concerned.[8] In all cases we present the number of tokens and types. It should be emphasized that these numbers have to be read relative to the number of possible distinctions present in the parsed corpus, and are not meaningful as absolute measures. For example, many of the PoS differences refer to subcategorization, and many people would argue against calling them PoS ambiguity. Still, this is the way the corpora were encoded, so it is at least one possible way to look at the matter.

Note that, if one knew that all corpora had been parsed by the same (version of the) parser, one could increase the number of ambiguous forms by adding up all possible analyses across corpora. That is, unique occurrences in one corpus could be identified as ambiguous with the help of occurrences in other corpora. We have not done this here, though.

Also, note that we have not taken into consideration the word forms analysed as belonging to a proper noun (named entity), each of which individually carries a PROP tag. So, we deleted them prior to inspecting potential ambiguity, as well as merged capitalized and non-capitalized forms in the computations above.

3.3. How to measure quality?

Ideally, one would pick all ambiguous forms and check them, in the way illustrated above – but this procedure would be as costly as to parse the whole corpus manually once again. So, the most obvious solution is to randomly select a subset of the ambiguous forms, and measure them, extrapolating as far as quality in the whole corpus is concerned.

We have thus randomly selected 100 cases (distinct types) of each kind of disambiguation, and analysed them. We have only taken into account non-capitalized words, in order not to add the additional question of recognizing proper names (named entities).

Corpus	PoS	Lemma	Morphology
ENPCANOT	11% 3%	12% 23%	3%, 25% 1%
EBRANOT	8% 4%	27% 7%	4%,16% 4%
NATPANOT	17% 1%	53% 6%	13%, 15% 9%

Table 6: Evaluation of 100 cases

The results appear in Table 6, presenting the percentage of analyses considered respectively wrong and about which there were doubts about how to classify it. For morphology, the intermediate number concerns the

[6] Except when followed by a non-finite verb, where the noun is parsed as subject of the following infinitive or gerundive clause.

[7] One could as well have a frequency estimate of the relative probability of each PoS, by itself or as an n-gram, etc.

[8] This means that they were assigned the same PoS. Lemma atribution and morphological marking were assessed independently.

forms which had only partly disambiguated information (considered as neither wrong nor doubtful). It should be noted that the lemma evaluation displayed in Table 6 reflects very often spelling errors, foreign words, and wrong PoS assignment. This is especially true for NATPANOT, where 42% of the cases inspected (and considered wrong) were due to errors in the original corpus text.

4. Automatic extraction of semantic relations from syntactic relations

We concentrate now on a specific task that uses parsed corpora as data for achieving a more complex goal. We apply a technique for automatic extraction of semantic relations from syntactic relations proposed in Gasperin (2001) and Gasperin et al. (2001), as an extended version of the technique proposed by Grefenstette (1994). This technique is based on the computation of word similarity through the syntactic contexts they share. (As syntactic context, we understand any word that establishes a syntactic relation with a given word in the corpus.)

We consider the following syntactic relations: an adjective as noun modifier, a noun as noun modifier (through a preposition), a noun as verb subject, a noun as verb direct object, and a noun as verb indirect object. The technique consists on extracting the syntactic contexts of each word from every occurrence of it in a parsed corpus, the words are compared as to occurrence in syntactic contexts, and words with many common syntactic contexts are considered semantically related. To perform the comparison, the similarity measure used is a weighted version of the Jaccard measure, that assigns global and local weights for each syntactic context. We then extract lists of semantically related words for each word in the corpus, which are useful mainly for thesauri construction.

The parsed corpus is thus necessary to extract the syntactic relations used in the procedure described above. We wanted to observe how dependent was the whole procedure on the correctness of the parsing information (specifically, PoS tags and function tags). In other words, if one extracts "wrong" syntactic contexts, how much this is reflected in the generation of noisy lists of semantically related words.

So, we present, on the one hand, measures of the robustness of the extraction of each syntactic relation used, and then some experiments about its influence on the semantically related words obtained as the result.

4.1. Measuring the extraction procedure

To measure the correctness of the syntactic contexts extracted from the corpus, it was necessary to compare them manually with the original expressions in the corpus, aiming to discover parsing problems. So, we adopted the following procedure:

1. selecting a portion of the FOLHANOT corpus;
2. selecting the nouns of this portion;
3. extracting all the syntactic contexts of these nouns;
4. comparing manually the extracted contexts with the original expressions in the corpus;
5. classifying the parsing performance.

The portion extracted from the FOLHANOT corpus contains around 5,000 words, where around 1,000 are nouns. The syntactic contexts of these nouns were extracted, some examples are shown on Table 7.

Sentence	Nouns	Contexts
... inicia a colheita da maior safra de sua história ... (... begins the crop of the largest production of its history ...)	*colheita*	<direct object, *iniciar*> <modifier, *de*, *safra*>
	safra	<adjective, *grande*> <modifies, *de*, *colheita*> <modifier, *de*, *história*>
	história	<modifies, *de*, *safra*>

Table 7: Examples of syntactic contexts

We classified each syntactic context extracted as: (C) correctly parsed, (E) incorrectly parsed, and (FE) it wasn't extracted due to a parsing error. Table 8 shows the percentages of the contexts according to these classes.

Class	Percentage (%)
C	89.96
E	7.82
FE	2.20

Table 8: Contexts according to parsing performance

Some erroneous contexts were more frequent than the others. The E and FE contexts were distinguished according to specific points. We can identify regular errors in the parsing information. Table 9 shows the most frequent parsing errors (or, in some cases, features) that generated the erroneous contexts, their percentage of occurrence and some examples.

Errors	Occurrence%	Examples
Proper nouns as common nouns	17.28	"Barreiras" (organization name) was treated as the common noun meaning barrier or barricade; "Folha" (newspaper name) was treated as the common noun meaning leaf
Prepositional attachment errors	14.81	"expansão de soja na fronteira" (soy expansion on the boundary): "fronteira" is attached to "soy" but should be attached to "expansão"
verb "haver" (in the form "há") as preposition	2.46	"instaladas no local há anos" (installed in the place for years)
preposition "a" as determiner and vice-versa	1.23	"se destina a implantação" (it is destinated to the implantation)
prepositional phrase as adverbial phrase and vice-versa	7.04	"disputar o campeonato na Holanda" (dispute the championship on The Netherlands): "na Holanda" should be as adverbial phrase

incorrect subject, direct object or indirect object tags	29.62	“impediu o plantio de feijão” (prohibited the plantation of beans): “de feijão” should be a prepositional phrase instead of an indirect verb object
adjective as verb	11.11	“ano passado” (last year): “passado” should be tagged as adjective instead of a verb form of to pass; “pesquisas confiáveis” (reliable research): “confiáveis” should be an adjective, not the verb “to rely on”
adjective as noun and vice-versa	7.40	“quinta": referring to “quinta-feira” (Thursday) instead of the ordinal number “quinto” (fifth); “alta de preço” (price increase): “alta” referring to “the increase” instead of the adjective “tall”
verb as noun and vice-versa	6.17	“corrida” (run): running event instead of the running action

Table 9: Most frequent parsing errors

It should be noted that, while from the point of view of the user (the extractor of syntactic contexts), they are considered errors, often the problems reported in Table 9 concern actual linguistic decisions made in the parsing process. For example, the classification of *há* as a preposition was an actual choice of the parser developer. The same happens with the PoS marking of past participles as verbs, not matter whether they are adjectivally used or not. Finally, even properties of the CG formalism, namely the underspecification of attachment, can be felt as problems and give rise to errors. This shows clearly, in our view, the different assessment types when one is involved in user-visible and not user-transparent evaluation.

After investigating the syntactic contexts, we used them to extract the semantic relations among the nouns.

4.2. Extracting semantic relations

To verify the influence of the erroneous syntactic contexts extracted from the corpus, we did two experiments: (1) we generated the lists of semantically related words to each noun using all the extracted contexts, and (2) we did the same using only the C and FE syntactic contexts. There is not a systematic measure to evaluate the homogeneity of the generated lists, so they were compared subjectively.

Table 10 presents the lists of semantically related words to some of the nouns in the corpus for both experiments.

To have a good homogeneity level, the used portion of the corpus should be larger. But in this paper we focus on the differences between the lists generated on each experiment, while expecting to report the results of a larger-scale experiment further in Gasperin et al. (in preparation).

We can observe that the lists corresponding to experiment 2 are more homogeneous than the lists produced by experiment 1. They are smaller and less noisy. The position of the words in the list indicates more or less similarity with the word in focus.

	Experiment	Semantically related words
expansão (expansion)	1	grosso exemplo lavoura monocultura t ha colheita
	2	lavoura monocultura t ha colheita
ha (hectare)	1	milho palanque monocultura quilo nelore t grosso
	2	quilo t km2 tonelada expansão

Table 10: Semantically related words in the two experiments

5. Conclusions

No matter the obvious usefulness of having parsed corpora available on the Web for interrogation, or as raw data for further NLP processing, the linguistic information carried by AC/DC corpora is still far from reliable in many cases. This is one of the reasons the Floresta Sintá(c)tica project was launched (Afonso et al., 2002a, 2002b), so that human revision could create more reliable resources.

For the majority of the readers of the present paper, though, who are not interested in Portuguese NLP in itself, we suggest the following general conclusions:

- one has to measure carefully what is the *difficulty* of a particular task, before trying to evaluate the result of performing that task
- there are implementable ways of measuring such a apriori difficulty, given a parsed corpus
- many apparently straightforward tasks, such as assigning objects or identifying tense or PoS turn out to be trickier than expected
- different applications and users may be interested in different properties and aspects of a parsed corpus, so one should evaluate *relative* to a given need.

6. Acknowledgements

We are most grateful to Vera Strube de Lúcia for her supervision of the second author in her dissertation on extracting semantic relations from syntactic contexts, without which this paper could not have been written.

7. References

Afonso, Susana, Eckhard Bick, Renato Haber and Diana Santos. (2002a). Floresta sintá(c)tica: um treebank para o português. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*. Lisboa: APL.

Afonso, Susana, Eckhard Bick, Renato Haber and Diana Santos. (2002b). Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC2002* (this volume).

Black, E., S. Abney, D. Flickinger, C. Gdaniek, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini and T. Strzalkowski. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop* (pp. 306--311).

Bick, Eckhard. (1998). Structural lexical heuristics in the automatic analysis of Portuguese. In Bente Maegaard (Ed.), *Proceedings of the 11th Nordic Conference on Computational Linguistics, Nodalida '98* (pp. 44--56). Copenhaguen.

Bick, Eckhard. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Carroll, John, Ted Briscoe and Antonio Sanfilippo. (1998). Parser evaluation: a Survey and a New Proposal. In Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (Eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Vol 1, pp. 447--454). Granada: ELRA.

Gaizauskas, Robert. (1998). Evaluation in language and speech technology. *Computer Speech and Language,* 12(4), 249-62.

Gasperin, Caroline. (2001). *Extração automática de relações semânticas a partir de relações sintáticas* [Automatic extraction of semantic relations from syntactic relations]. MSc thesis, Porto Alegre, Brazil: PPGCC-PUCRS.

Gasperin, Caroline, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes and Vera de Lima. (2001). Using Syntactic Contexts for Measuring Word Similarity. In Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli (Eds.), *Proceedings of the workshop "The Acquisition and Representation of Word Meaning", ESSLI'01*. Helsinki.

Gasperin, Caroline, Diana Santos and Vera Strube de Lima. (In preparation). Semantic relatedness among words: what is required from syntax?.

Grefenstette, Gregory. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Hindle, Donald and Mats Rooth. (1993). Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1), 103--120.

Johansson, Stig, Jarle Ebeling and Signe Oksefjell. (1999). English-Norwegian Parallel Corpus: Manual. Univ. of Oslo: Department of British and American Studies, http://www.hf.uio.no/iba/prosjekt/ENPCmanual.html

Lin, Dekang. (1995). A dependency-based method for evaluation broad-coverage parsers. *Proceedings of IJCAI'95* (pp. 1420--1425). San Mateo, Calif: Morgan Kaufmann Publishers.

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

Reis, Regina. (1993). Dicionários de língua corrente: algumas considerações. In *Actas do 1.o Encontro de Processamento de Língua Portuguesa (Escrita e Falada), EPLP'93* (pp. 141--146). Lisboa.

Santos, Diana. (1996). Português Computacional. In Inês Duarte and Isabel Leiria (Eds.), *Actas do Congresso Internacional sobre o português* (Volume III, pp.167--184). Lisboa: Edições Colibri / APL.

Santos, Diana. (1999). Toward Language-specific Applications. *Machine Translation,* 14(2), 83--112.

Santos, Diana and Eckhard Bick. (2000). Providing Internet access to Portuguese corpora: the AC/DC project. In M. Gavriladou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (pp. 205--210). Athens: ELRA.

Santos, Diana and Signe Oksefjell. (1999). Using a Parallel Corpus to Validate Independent Claims. *Languages in contrast*, 2(1), 117--132.

Santos, Diana and Paulo Rocha. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 442--449). ACL.