Grammatical annotation of the Portuguese C-ORAL Corpus

Eckhard Bick
(University of Southern Denmark)

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark
Campusvej 55
DK 5230 Odense M
Denmark
e-mail: eckhard.bick@mail.dk
web http://visl.sdu.dk

1. Introduction

Over the last two decades, Corpus Linguistics has experienced a gradual widening of its research focus to encompass not only written, but also spoken language data. Both in scope and volume this shift has been quite successful, and today hundreds of spoken language corpora are available. However, availability is often restricted to individual research groups, or hampered by the lack of systematic searchability. This second limitation is closely linked to annotational issues - in order to be optimally useful, a speech corpus should have a fine-grained and standardized annotation at various levels, such as prosody, discourse structure etc, but also traditional morphosyntactic annotation. In this paper we will focus on how to integrate the latter with the former, and discuss the question whether and how a tagger-parser primarily designed for written language can be adapted to handle transcribed speech data. The work was carried out in the research context of the C-ORAL speech corpus project for Brazilian Portuguese (Raso & Mello 2010), where morphosyntactic annotation was to be added automatically on top of an existing meta-annotation in the face of non-standard orthography and the absence of punctuation, preserving in-text speech flow markers etc.

Using automatic annotation, either on its own or as a pre-step for manual revision, is an obvious choice for a corpus this size (~ 300.000 words). Thus, previous European C-ORAL sister projects employed statistical part of speech taggers for this task, such as the PiTagger system (Moneglia et al 204) for the Italian section, which had access to a lexicon-based analyzer, a standard lexicon (107.00 lemmas), a training corpus (50.000 words) and a special pre-dictionary covering about 2000 non-standard and dialectal forms. For the European Portuguese section, the Brill tagger (Brill 1993) was used, trained on a written Portuguese corpus of 250.000 words. While no higher-level, syntactic annotation was attempted in the European C-ORAL, other speech corpus projects have opted for full treebank annotation, such as the Arabic treebank describe by Maamouri et al. (2010), which combined manual selection of analyzer suggestion, followed by an automatic syntactic parsing stage. However, the Arabic treebank was built from broadcast data, not interviews or spontaneous dialogue, so no direct comparison can be made with C-ORAL, given the much lower need for non-standard forms and discourse meta annotation in the transcription of news feed data.

For our own work we used the Palavras parser (Bick 2000) as a point of departure. Palavras is a Constraint Grammar (CG) parser that is mostly used for the annotation of written data, but has demonstrated great robustness in the face of genre variation (as, for instance, in the Linguateca[1] project and the CorpusEye corpora[2]). With lexical adaptation and various filter programs, the parser has also been used for non-standard language varieties, such as historical texts (Bick & Módolo 2005). The Constraint Grammar paradigm (Karlsson 1995), which the Palavras parser adheres to, can be described as a dualism of a robust, modular disambiguation methodology for Natural Language Processing (NLP) on the one hand, and a linguistic-descriptive convention on the other hand, encoding linguistic analyses as token-based tags and function-mediated dependency structures. Both the method and the descriptive tradition offer a number of formal advantages for the annotation of non-standard language data such as speech. First, because CG systems have a modular architecture with a clear separation of lexica, analyzers and grammars (rule sets) for successive levels of analysis, it is relatively easy to add specialized lexica or morphological filters, as well as add specific grammar modules. Second, CG's token-based annotation, where even higher-level structural information is strictly token-based, allows a corpus project to maintain several layers of annotation in parallel (such as discourse markers as opposed to clause boundaries), even allowing rules handling one layer to make reference to tags from another layer. Several speech annotation projects have made use of these advantages, such as Müürisep & Uibo (2006) for

[1] www.linguateca.pt
[2] www.corp.hum.sdu.dk

Estonian and Bondi et al. (2009) for the Nordic Dialect Corpus, though the latter used a hybrid technique, where written-text CG was used to annotate a chunk of speech data from the Oslo area, which was then manually corrected and used to train a Decision Tree Tagger (Schmid 1994) for use on other (Norwegian) dialects. In the European C-ORAL context, the Spanish section employed CG-inspired rules for part-of-speech disambiguation of morphological output from the GRAMPAL system (Moreno 2003), and for the Palavras parser itself, Bick (1998) reports early experiments with a Constraint-Grammar-only solution in connection with the morphosyntactic annotation of the Brazilian NURC corpus ("Norma Lingüística Urbana Culta", Castilho 1993).

2. The point of departure: The Palavras Constraint Grammar parser

Technically, the Palavras parser is a chain of Constraint Grammar rule sets, successively handling ever higher (deeper) levels of analysis, progressing from morphological disambiguation and PoS tagging, over syntactic function mapping and dependency relations, to semantic role annotation, Named Entity Recognition and application-oriented modules. Input to this chain of grammars is provided by a preprocessor/tokenizer and a morphological analyzer program supported by large lexica covering inflexional paradigms, valency potential, semantic class ontologies etc. All lexical information is encoded, CG-style, as token-linked tags on reading lines. Ambiguous reading lines for a given word are called a *cohort,* as would be the case for the typical verbo-nominal Portuguese ambiguities involving -a and -o endings*:*

"<casa>"
"casa" <build> N F S ('house')
"casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V IMP 2S VFIN ('marry!')
"casar" <vt> <vi> <com^vp> <vr> <com^vrp> <vH> V PR 3S IND VFIN ('marries')
"<accordo ALT acordo>"
"acordo" <sem-c> <+com> <+sobre> <+entre> <de+> <+n> ('deal, contract')
"acordar" <ve> <vt> <vK> V PR 1S IND VFIN ('wake up')

A distinction is made between primary tags, which are slated for disambiguation (e.g. N, V), and secondary tags that are not (or not at this level) intended for disambiguation (<...> tags), but rather to provide contextual clues for CG rules in the process of primary disambiguation. Thus, a transitive verb tag <vt> and a human noun tag <H> may help to assign subject and object functions to the nouns in a sentence.

Annotated output from Palavras will optimally provide only one primary tag[3] of each kind (though tag strings may be complex, as in the case of a morphological gender-number tag string):

O <artd>	DET M S	@>N	#1->3
último	ADJ M S	@>N	#2->3
diagnóstico	N M S	@SUBJ>	#3->9
elaborado	V PCP2 M S	@IMV @#ICL-N<	#4->3
por	PRP	@<PASS	#5->4
a <artd>	DET F S	@>N	#6->7
Comissão=Nacional	PROP F S	@P<	#7->5
não	ADV	@ADVL>	#8->9
deixa	V PR 3S	@FMV	#9->0
dúvidas	N F P	@<ACC	#10->9
$.			#11->0

[3] For subclauses, two notational conventions can be chosen. While the VISL-compatible, newer standard encodes subclause function as a single tag on the head verb (first verb) of the subclause, the original PALAVRAS convention uses 2 tags for subclauses, on either the subordinator word (finite subclauses) or the verb (non-finite subclauses), one tag being internal (@), and one external (@#) for the function of the subclause as a whole.

The examples shows an annotation at the tree structure level, with tag fields for part of speech (N=noun, ADJ=adjective, V=verb etc.), morphology (M=male, F=female, S=singular, P=plural etc.) and syntactic function (@SUBJ=subject, @>N=prenominal, @#ICL-N< = relative, postnominal non-finite clause, @<ACC =accusative object, @P< argument of preposition, @FMV=finite main verb, @IMV=non-finite main verb). The dependency links are already implicit at the syntactic function level, through > and < attachment direction markers, pointing towards the head of the phrase or clause. The complete dependency tree, finally, is provided by #n->m "arc tags", where n is the token ID, and m the ID of its dependency mother. Syntactic tree structures like these can be transformed into a variety of formats. For complete trees, PALAVRAS offers, for instance, traditional Chomskyan constituent brackets, PENN treebank annotation, TIGER treebank xml and MALT dependency xml markup.

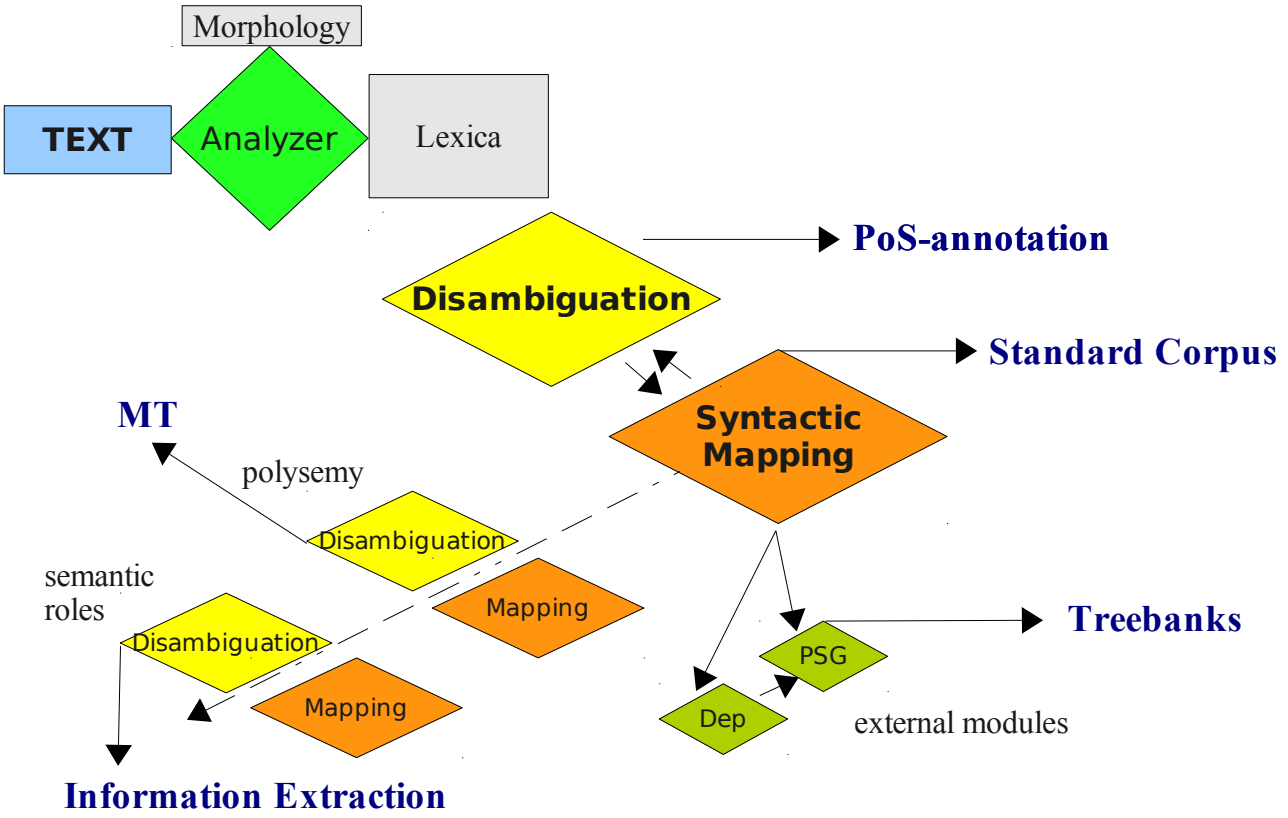

Fig. 1: Parser flow chart

PALAVRAS uses about 6.000 contextual CG rules that either remove, select, add, map or substitute tags/readings. Apart from other tags (associated with any other word in the sentence), the CG formalism allows rules to make reference to word- or reading-related statistical-numerical information, match regular expressions in word forms, tags and lexemes, or unify category features across constituents. Though somewhat foreign to the formalism's reductionist methodology, even generative rewriting rules can be expressed in Constraint Grammar, with the help of so-called templates.

Most rules, however, are fairly straight forward disambiguation rules such as the following, that removes a finite verb reading (VFIN), if there is a safe (C) preposition reading to the left:

REMOVE VFIN IF (-1C PRP) ; # remove a

While such close context could be captured by probabilistic Hidden Markov Model (HMM) n-gram modeling, too, many CG rules have global sentence scope and are considerably more powerful. Consider for instance the following uniqueness rule, saying that a word cannot be a finite verb if

there already is a finite verb anywhere (*) to the left (*-1), without a clause boundary (CLB) or coordinator (KC) in between (BARRIER).

REMOVE VFIN (*-1 VFIN BARRIER CLB OR KC)

Given the architecture and rule methodology of the parser, three challenges can be identified with regard to its application to oral data, affecting lexical recall on the one hand (2) and contextual disambiguation on the other (1,3). In many ways, the problems are similar to the ones encountered in the annotation of historical language data (Bick & Módolo 2005).

1. How to maintain corpus meta information from the non-grammatical annotation layers, while still providing "running text" input to the parser and its analyzer
2. How to adapt the lexicon and/or to change the word forms to allow input to be recognized as ordinary written, modern, standard Brazilian Portuguese, while at the same time maintaining the oral transcription forms
3. How to provide syntactic breaks, in the absence of ordinary punctuation, to allow the definition of delimited windows for contextual disambiguation

We will treat these issues in chapters 3, 4 and 5, respectively.

3. Text flow normalization

C-ORAL-Brasil uses a number of symbols and encoding conventions to handle data flow issues like turn taking, prosodic breaks, speaker overlap, retractions and interruptions. Such encoding is either in non-alphanumeric form (<, /, +), or not part of an utterance (speaker names), so they either cannot or must not be analyzed by the parser. To both maintain this meta-information and to provide text-only input to the parser, we opted for a two-level annotation, where meta-information is "stored" in angle brackets on separate lines as corpus meta markup, reminiscent of e.g. <source>, <s> and <p> markers in written corpus annotation. PALAVRAS' annotation is transparent to such markup and will not change, remove or try to analyze it. Consider the following two-turn example, first in C-ORAL native annotation, then in vertical CG format, after parsing.

***LEO:** o Juninho **<foi>** //
***GIL: <ô / mas>** / voltando à questão / **falando em [/2]** e também falando em povo mascarado / esse povo do Galáticos é muito palha / eu acho que es nũ deviam mais participar / e **<tal>** //

<LEO:>
o [o] <artd> DET M S @>N
Juninho [Juninho] <hum> <newlex> <*> PROP M S @SUBJ>
<overlap-start>
foi [ser] <fmc> V PS 3S IND VFIN @FMV
<overlap-stop>
$;
<GIL:>
<overlap-start>
ô [ô] <newlex> IN @ADVL
$,
mas [mas] KC
<overlap-stop>
$,
voltando [voltar] V GER @IMV @#ICL-ADVL>
a [a] <sam-> PRP @<PIV
a [o] <-sam> <artd> DET F S @>N
questão [questão] <ac> N F S @P<
$,
<retract:falando_em>
e [e] KC
também [também] ADV @ADVL>
falando [falar] <vH> V GER @IMV @#ICL-<ADVL
em [em] PRP @<PIV
povo [povo] <HH> N M S @P<
mascarado [mascarar] <vH> V PCP M S @N<
$,
esse [esse] <dem> DET M S @>N
povo [povo] <HH> N M S @SUBJ>
de [de] <sam-> PRP @N<
o [o] <-sam> <artd> DET M S @>N
Galáticos [Galáticos] <org> <newlex> <*> PROP M P @P<
é [ser] <vK> <fmc> V PR 3S IND VFIN @FMV
muito [muito] <quant> ADV @<ADVL
palha [palha] <cm> N F S @<SC
$,
eu [eu] PERS M/F 1S NOM @SUBJ>
acho [achar] <vH> <fmc> V PR 1S IND VFIN @FMV
que [que] KS @SUB @#FS-<ACC
es OALT eles [eles] PERS M 3P NOM @SUBJ>
nũ OALT não [não] ADV @<ADVL

```
deviam   [dever] V IMPF 3P IND VFIN @FAUX
mais     [mais] ADV @<ADVL
participar          [participar] <vH> V INF @IMV
     @#ICL-AUX<
$,
e     [e] KC
<overlap-start>
tal   [tal] <diff> <KOMP> DET M/F S @<OC
<overlap-stop>
$;
```

Here, only lines not starting in '<' are part of the morphosyntactic annotation. Speaker names are separate meta tags <GIL:>, and overlaps (<....>) are marked with <overlap-start> and <overlap-stop> markers. It is a clear advantage for the parser that retractions are pre-marked manually in brackets at the start point of the retraction, providing the precise number of retracted words. Our preprocessor module only needs to eliminate the words in question from the surface level to enable much smoother syntactic parses. Word repetitions or self-corrections, if allowed to persist at the surface level, would be problematic for CG rules at all levels, interfering not only with the implementation of linguistic universals like the uniqueness principle, but also with word class adjacency and agreement rules.

As can be seen from the example, the surface-deleted words will be stored in a special <retract:...> tag, maintaining the principle of two-level annotation, where two levels of annotation are separated, but not mutually exclusive. The same procedure is used for so-called non-words, which come in 2 types - first, a few non-word surface strings without special markup ('hhh' and 'xxx'), and second, incomplete words (contractions), which are marked with an initial &-sign.

*GIL: **hhh** eu tenho **&dire**

```
<GIL:>
<nonword:hhh>
eu          [eu] PERS M/F 1S NOM @SUBJ>
tenho       [ter] <fmc> V PR 1S IND VFIN @FMV
<nonword:&dire>
```

Since the above rewriting of surface markings as corpus meta tags needs to access running text rather than individual words, and is meant override PALAVRAS' own tokenization, it has to be done by a preprocessor prior to tokenization. In terms of module succession, this is also the place where PALAVRAS multi-word treatment can be overridden. so multi-word expressions (MWEs) from corpus specific lexicon files are used to create “multi-words” such as *'empepê=três' (MP3), 'al=dente', 'air=bags',* preventing PALAVRAS from assigning partial analyses, and instead providing the corpus specific tagging lexicon with appropriate forms. Thus, *'air=bags'* will not receive a native PALAVRAS analysis (default noun singular), but rather a noun plural reading from our corpus lexicon (N M P).

A special complication arose from the fact that overlap and retraction markings can be nested and/or overlapping, as the example below shows, requiring careful ordering of string matches, for instance to prevent retractions from getting “invisible” within (de-texted) speaker overlap markers. Also, since overlaps and non-words can appear within the scope of a retraction, they would change the latter's word count if removed too early, and possibly affect real words further to the left.

*GIL: <eu &a [/2] eu acho que é> esse [/2] é esse aqui o' // <&he> +

```
 <GIL:>
<overlap-start>
<retract:eu_&a>
eu    [eu] PERS M/F 1S NOM @SUBJ>
acho [achar] <vH> V PR 1S IND VFIN @FMV
que  [que] KS @SUB @#FS-<ACC
<retract:é>_esse>
é     [ser] <vK> V PR 3S IND VFIN @FMV
esse       [esse] <dem> DET M S @<SC
aqui       [aqui] ADV @N<
o' OALT olha [olhar] <vH> V PR 3S IND VFIN @FMV
$;
<overlap-start>
<nonword:&he>
<overlap-stop>
$...
```

It should be noted that non-inclusive bracketing overlaps of the type <a> represent a general annotation problem, even for elaborate xml encoding schemes, since the latter do not envision non-projective (overlapping) tree structures, so the CG annotation chosen here can be said to be a fairly robust solution.
Portuguese, especially in spoken language, employs a special focus construction with *ser .... que,* which in some cases formally lends itself to a syntactic analysis with an absolute relative clause (*o que),* but more often than not, usage support a simpler, more functional analysis with *e=que, foi=que* or simply *é* as a focus particle, inserted before the focused constituent:

***é** uma cerveja **que** quero*
*--> uma cerveja **é que** quero*
*-> quero **é** uma cerveja*

In the C-ORAL-Brasil corpus, the focus particle é=*que,* occurs 380 times, in ~ 2% of turns, but is transcribed as *que,* Since PALAVRAS would read an ordinary *que* as a conjunction or relative, rules would run into difficulties due to the absence of a subordinate clause. Therefore, the normalization preprocessor tries to match sequences of *qu*-words and *que* (que que, quando que, quanto que, quem que, onde que), and insert the standard *e=que,* while retaining the substituted sequence in an <elision:....> meta tag:

*LEO: <beleza> // <então a gente já sabe em quem que a gente vai colocar a> culpa //

```
<LEO:>
<overlap-start>
beleza     [beleza] <am> N F S @NPHR
<overlap-stop>
$;
<overlap-start>
então      [então] <kc> ADV @ADVL>
a     [o] <artd> DET F S @>N
gente      [gente] <HH> N F S @SUBJ>
já    [já] ADV @ADVL>
sabe       [saber] <fmc> V PR 3S IND VFIN @FMV
em  [em] PRP @ADVO>
<ellision:quem_que>
quem       [quem] <interr> SPEC M/F S/P @P<
é=que      [é=que] <foc> ADV @<FOC
a     [o] <artd> DET F S @>N
gente      [gente] <HH> N F S @SUBJ>
vai  [ir] V PR 3S IND VFIN @FAUX
colocar    [colocar] <vH> V INF @IMV @#ICL-AUX<
a     [o] <artd> DET F S @>N
<overlap-stop>
culpa      [culpa] <am> N F S @<ACC
$;
```

Since PALAVRAS is not just a part of speech tagger, but a syntactic parser providing deep tree analyses, it needs prepositions and pronouns to fill their respective syntactic slots (in pp's and np's) even in the case of surface form contractions such as *deles (de eles), num (em um), pelos (por os).* The resolution of contractions makes syntactic structure more transparent and makes linguistic generalisation easier. Thus, it has advantages both for the disambiguation grammar (verbal valency can "see" a given preposition, nouns can "see" their article) and facilitates tasks like np-extraction, cross-language word alignment and the extraction of collocation patterns in lexicography. However, while the parser automatically splits contractions with the prepositions *de (~52%) , em (~37%), a (2.6%)* and *por (1.7%),* as well the historical forms with *com (2%)* and some frequent contractions with *para (5%),* it did not cover all combinations with the latter, and obviously missed out on contractions with non-standard second parts, such as *naquea (em aquela).* Therefore, the remaining forms had to be expanded by the C-ORAL-preprocessing, either (a) before or (b) after PALAVRAS' tokenization step. For pre-tokenization, a regular expression-match was used, and a <contraction:...> meta-tag inserted in front of the expansion, which all were 2-part expansions, given below with their corpus frequencies:

pa (477), pro (122), co (23), pros (20), prum (18), pos (18), ca (15), pras (14), cum (9), cos (7), des (8), cos (7), pum (6), puma (5), naquea (5), cuma (5), pruma (4), dea (3), daquea (3), pas (1), naques (1), daqueas (1).

eu falei isso naquea reunião lá

```
eu   [eu] PERS M/F 1S NOM @SUBJ>
falei      [falar] <vH> V PS 1S IND VFIN @FMV
isso       [isso] <dem> SPEC M S @<ACC
<contraction:naquea>
em  [em] PRP @<ADVL
aquela     [aquele] <dem> DET F S @>N
reunião    [reunião] <occ> N F S @P<
lá     [lá] ADV @N<
```

The most problematical case was *pra,* because this form is ambiguous - it can either be a shortened version of *para,* or a two-word contraction, *para a,* calling for contextual disambiguation.

/ só **pra** eles mesmos // *(=para)*
// **pra** próxima taça / *(=para a)*

Post-tokenization *(coral.inter*-program*)* was used for the contractions that were less regular and/or more difficult to match with regular expressions. These cases were drawn from C-ORAL's normalisation lexicon, and their parts were word-form numbered and marked with OALT normalization tag (cp. chapter 4), in theory allowing any number of parts:

*pa despesa é bastante / **né** //*

```
pa OALT pra       [para] <sam-> PRP @ADVL>
a          [a] <artd> <-sam> DET F S @>N
despesa    [despesa] <mon> N F S @P<
é          [ser] <vK> V PR 3S IND VFIN @FMV
bastante   [bastante] <nh> ADJ M/F S @<SC
$,
<slash>
né OALT não       [não] ADV @ADVL>
né-2 OALT é       [ser] V PR 3S IND VFIN @FMV
$;
```

Note the standardization of *pa* as *pra,* not *para,* leaving it to PALAVRAS to assign and resolve the *para=a/para* ambiguity.

4. Lexical and orthographic normalization

In order to assign a morphological tag string and word class hypothesis, PALAVRAS tries to recognize unknown words as either (1) affix-derivations or (2) variations of standard forms, or a combination of both. Even for written language data, (2) is an important robustness factor because of the spelling differences between European and Brazilian Portuguese (a), oi-ou and accent variation (b) etc., as well as the need to understand texts with a spelling that has been rendered obsolete by orthographic reform. Even some typos and historical forms are handled this way. The recognized standard form will be juxtaposed to the original word form as an ALT tag:

(a) dicção ALT dição [dição] <sem-s> N F S
(b) négócio ALT negócio [negócio] <act-d> N M S

This way, the full tag type scheme for the C-ORAL-Brasil annotation will look like this:

wordform (ALT normalization) [lemma] <secondary tags> PoS MORPHOLOGY @SYNTAX

In some rare cases, the analyzer will change an unknown, but existing word in order to match a derivational analysis, e.g. read *hemácia* as *hemacia (hem-ac-ia).* While leading to a lexeme error, this method will still in most cases yield a correct PoS and morphological analysis.

For the C-ORAL project, however, ordinary standardization was deemed not to be enough, first of all because certain oral word forms were transcribed in a phonetic fashion *as is*[4], creating in some

[4] Transcription of oral data has to strike a balance between standardization and phonetic fidelity. Too little standardization will make the corpus difficult to use and search, to do lexical frequency analysis or word order studies. Too little phonetic fidelity, on the other hand, will remove some of the very features and patterns we might want to learn from the corpus. Thus, a question like "how common is '-im' as a diminutive?" or "how common is s-drop in verbal inflexion?" can obviously not be answered if full normalization is used. Therefore, only two level annotation like

cases unrecoverable differences from standard orthography, or the risk of ambiguity. As a side consideration, we also wanted to account for lexical gaps due to dialectal or otherwise rare forms. Therefore, two new modules were added to PALAVRAS' program chain, both with a manually maintained lexicon-file as input. The first program *(coral.inter)* handles specific or systematic standardizations and is run after preprocessing, before morphological analysis, while the second program *(postlex_pt)* is regular morphological analyzer in its own right, with its own lexicon and inflexion rules, overriding PALAVRAS' own analysis, removing the error risk created by heuristic readings.

An example for systematic normalization is the addition of first person plural -s for verbs *(comemoramo -> comemoramos, encontramo -> encontramos)*, which *coral.inter* accomplishes using string matches and a fullform lexicon that helps to avoid false s-additions to e.g. nouns like *balsamo, dinamo, esperramo.* l-r variation *(glandão - grandão)* was also covered but proved to be negligible in quantitative terms.

About 700 normalizations were listed in a special lexicon file[5], and though the standard analyzer could have handled a certain proportion on its own in terms of word class, the lexicon treatment also allowed us to add correct base forms or even semantic classification. A very phonetic example are abbreviations (a1-3) where even plural (a2) forms and non-standard pronunciation (a3) were covered[6]. Other groups cover non-standard inflexion (d1-3) and derivation (c1-2). Finally, word-initial changes like a-drop (b2-4) had to be covered in order to prevent such forms from being guessed as (most likely) singular nouns.

(a1)	emedebê	MDB
(a2)	emeeles	ml
(a3)	emitivi	MTV
(b1)	envinha	vinha
(b2)	garrou	agarrou
(b3)	inda	ainda
(b4)	roz	arroz
(c1)	espim	espinhos
(c2)	ladim	ladinho
(d1)	estudemo	estudamos
(d2)	fazido	feito
(d3)	fize	fiz

While maintaining the original word form, standardized forms were added with an OALT:... prefix, and it is the standard form that annotation tags refer to:

meninim OALT menininho [menino] <DERS> N M S

The standardization lexicon also covers multi-word strings *(a'=aqui -> olha=aqui, c'=ocês -> com=vocês),* which is why the tokenizer preprocessor also needs access to the file. One advantage of multi-word normalization is that the individual parts provide disambiguation context for each other, allowing, for instance, the recognition of *a'* as *olha',* rather than the preposition or determiner reading, or the resolution of *n'* as *não* or *em* in *n'=era* and *n'=ocê,* respectively.

The second lexical add-on program, the override analyzer, is considerably more sophisticated than

---

the one we propose, allowing both form- and category-searches at the same time, may hope to combine the best of two worlds.

[5] Most of the original content for both the normalization file and add-on lexicon file was provided by one of the C-ORAL authors, Heliana Mello, followed by consistency and compatibility checks for the individual items, ensuring full tag coverage and preventing unwanted interferences with PALAVRAS' main lexicon.

[6] PALAVRAS does handle phonetic abbreviation spelling to, but only for base forms of type (a1), and by analyzing letters as "suffixes" (<DERS>): emedebê **"M"** <DERS **D**> <DERS **B**> b

the normalization program, and allows both fullform and base form entries in its lexicon *(newlex_pt)*. Regular inflexions of noun, adjective and verb forms will be recognized from the base form alone, but all irregular forms have to be entered separately. Like for the standardization lexicon, multi-word entries will also be visible to the preprocessor for tokenization (d1, b).

In the actual lexicon (currently 2000 entries), due to the good coverage of PALAVRAS, there are very few regular Portuguese nouns, and those there are could mostly have been recognized by PALAVRAS' derivational analysis (a1). Still, some inflected and complex forms (a2-3) may be useful to avoid the choice of a competing heuristic analysis, e.g. *caça-talentos* as plural- vs. singular-inflected. Also, the corpus contained a certain number of foreign words which are likely to be singular nouns, but may have endings that could trigger a heuristic (Portuguese) analysis as something else, e.g *remote* (c1). Even more important is it to list foreign non-noun words such as verbs (c3), adjectives (c4) or adverbs (c5), but these entries raise two problems that would have to be resolved if the lexicon were to be used in a more general setting (i.e. for other corpora): First, foreign words would need to be specified with *all* their readings, not only the one occurring in the corpus, e.g *shift* (c4) as both noun and verb. Second, also foreign entries would need full morphology, if they were to fully interact with their Portuguese context and CG-rules (e.g. agreement issues). This latter consideration has already been taken into account by semi-automatically adding singular male features (N M S) to noun entries without pre-entered morphology, but a similar strategy would be more difficult for verbs and adjectives due to the fact that English under-specifies adjective number and - in many forms - verb finity.

The largest portion of the lexicon, however, amounting to two thirds of all entries, are proper nouns (e1-3). Though these could be fairly safely recognized as such by PALAVRAS, their gender (and possibly number) is not easy to guess (e.g. *TIM* as feminine), and the addition of a semantic prototype reading (e.g. <hum>=human, <org>=organization, <Lciv>=town or state) provided valuable semantic context for CG rules, allowing, for instance, to unify the ±HUM feature on verbs and their subjects, allowing semantics-based disambiguation of word-class or syntactic function.

(a1) fazeção <activity> N F S
(a2) zenes N M P # termo de jogo
(a3) caça-talentos N M S
(a4) superbonitinha ADJ F S
(a5) superbem-arrumada ADJ F S
(b) mil-oitocentos-e=vovó=gostosa NUM M/F P
(c1) remote N M S # estrangeirismo
(c2) completed ADJ M/F S/P # estrangeirismo
(c3) save V # estrangeirismo
(c4) shift N M S # estrangeirismo
(c5) anche ADV # estrangeirismo
(d1) tu=tu X # onomatopéia
(d2) tuf X # onomatopéia
(e1) Titina <hum> PROP F S
(e2) TIM <org> PROP F S # operadora de telefonia
(e3) Timoftol <cm-rem> PROP M S
(f) agadê N M S # HD (harddisk)

In principle, the override lexicon module could be regarded as a general lexicon extension for PALAVRAS, since most specific adaptations, such as the afore-mentioned phonetically spelled abbreviations (f), or the female form of adjectives (a4-5), while not in standard format, do not disturb the morphological system of PALAVRAS either. On the other hand, the list contains a few entries with no regular word class, such as the onomatopoeia *tu tu* and *tuf* (d1-2), and the treatment of numerical expressions as wholes (b) is in conflict with the otherwise slightly more analytical approach used by PALAVRAS. Therefore, these word types, as well as the ambiguity potential of foreign words, should be consistency-checked before porting the lexicon to other corpora.

Originally, the C-ORAL lexicon extension was intended as a "hard override", i.e. the idea had been to use the provided analysis *instead of* the original PALAVRAS analysis, assuming the latter would be heuristically wrong or underspecified. However, introducing a new reading, inspired by corpus inspection, because PALAVRAS did not provide the desired analysis in a particular utterance, does not, for ambiguous words, necessarily mean that PALAVRAS would have come up with the wrong analysis every time (i.e. in other contexts). And since it is more difficult for a human to come up with a full cohort of ambiguous readings than for a computer (the brain simply filters out contextually meaningless alternatives), a more cautious solution had to be chosen, where the C-ORAL lexicon *adds* to PALAVRAS' own suggestions, rather than entirely replacing them. Since this is done *before* CG disambiguation, the grammatical rules then get a chance to choose the contextually best reading. At the same time, a bias was introduced that allowed the C-ORAL lexicon (marked <newlex>) to override PALAVRAS readings marked as heuristic[7] more easily than those supported by the PALAVRAS core lexicon and inflexional rules. An example of such unintended ambiguity interference is the word "pô", listed in the C-ORAL lexicon as an interjection: Because the general conventions of the C-ORAL-Brasil transcription allowed plural inflexions of interjections, this meant that the word form "pôs" would also be tagged as an interjection, competing with the common, verbal reading[8]. The example sentence shows, in rule-tracing mode, how the morphological CG module disambiguates the verb/interjection ambiguity in the sentence *pôs, ele pôs a mão na massa.* Discarded reading lines are prefixed with a semicolon, and the ID-numbers of the rules used are traced in bold face.

```
"<pôs>"
        "pô" <newlex> IN
;       "pôr" V PS 3S IND VFIN REMOVE:5712
"<$,>"
"<ele>"
        "ele" PERS M 3S NOM/PIV
;       "ele" N M S REMOVE:6237
"<pôs>"
        "pôr" V PS 3S IND VFIN  SELECT:5894
;       "pô" <newlex> IN SELECT:5894
"<a>"
        "o"  <dem> DET F S
;       "a"  PRP REMOVE:4756
;       "a" N M S REMOVE:4778
;       "ela" PERS F 3S ACC REMOVE:6729
"<mão>"
        "mão" N F S
"<em>"
        "em" <sam-> PRP
"<a>"
        "o" <-sam> <art> DET F S
"<massa>"
        "massa" N F S
"<$.>"
```

5. Syntactic segmentation

While written language data provide paragraph markers, line breaks, full stops and other punctuation to deduce syntactic and informational structure, such segmentation is implicit rather than explicit in spoken language transcriptions. At the surface level, speech data lacks punctuation and has, text-wise, unclear sentence and clause boundaries. To make things worse, speech data is filled with "syntactic" noise, such as repetitions, false starts and pauses or phatic interjections *(ah, eeh, uh),* However, the transcription-embedded information allowed us to overcome most of these problems and to turn the data into a textual format that can be processed by an ordinary parser.

Chapter 3 has described our solution for "syntactic noise", and we will now focus on segmentation. The necessary information to segment speech resides in prosody (i.e. rhythm, stress and intonation) as well as nonverbal signals. Depending on whether and how this information is encoded in the transcription, a parser may simply lack the segmentational information to work properly. Some speech corpora, such as the NURC corpus version described in (Bick 1998), use orthographic means to express vowel length (*'u::m'),* stress *('esnoBAR')* and even pauses *('eee'),* adding further word recognition difficulties, and the need for the insertion and contextual disambiguation of

[7] e.g. carrying <heur> or derivation tags (<DERS>, <DERP>), or lacking a frequency tag.

[8] and possibly the adverb *pois,* which however was not listed as *pôs* in the standardization lexicon.

pauses on the one side and true syntactic breaks on the other. In the C-ORAL corpus, on the other hand, rather than embedding prosodic information the orthography, prosodic segmentation was marked explicitly, at transcription time, using three different segmentation strengths:

1. major prosodic breaks (//), separating what functionally could be called utterances, equivalent to written language sentence separation.
2. discontinuation breaks (+) between utterances
3. non-terminal prosodic breaks (/), separating what could be viewed as informational units

Rather than making this information invisible to the parser by turning it into meta-tags (the strategy chosen for syntactic noise), we decided to replace the prosodic markers with standard punctuation, using a semicolon as the most obvious equivalent to the // terminal breaks (alternating with '...' for interruptions), and a comma for the non-terminal breaks (/). Portuguese orthography does not use obligatory commas in all places where our transcription had a slash, but inspection of annotation results showed that the extra commas helped rather than hurt. In CG-terms, a comma is a member of the BARRIER set in many context rules, separating phrase-internal material from tokens belonging to another phrase. It is therefore the global context rules (*1 and *-1 contexts) that stand to profit from the introduction of prosodic punctuation marks. In the same vein, syntax should be more affected than PoS/morphological tagging, since long-distance contexts are more important for capturing syntactic relations. A breakdown of rule scopes in the Palavras grammar (Bick 2000) illustrates the relative importance of global context for different tasks and heuristicity levels:

	morpological targets				**syntactic targets**				**all**
	safe	**heuristic levels**			**safe**	**heuristic levels**			
		1.	2.	3.		1.	2.	3.	
REMOVE tag (only local contexts)	403	112	13	27	153	37	4	2	651
REMOVE tag (≥ 1 global contexts)	183	44	5	5	941	219	17	1	1415
local/global	**2.2**	2.5	2.6	5.4	**0.2**	0.2	0.2	2.0	**0.5**
SELECT tag (only local contexts)	271	70	8	7	60	2	1	1	420
SELECT tag (≥ 1 global contexts)	129	23	9	2	209	57	3	-	432
local/global	**2.1**	3.0	0.9	3.5	**0.3**	0.0	0.3	-	**1.0**

The numbers show that the share of global rules is substantial even for morphology (around 31%), and is very high for syntax, where most rules use unbounded contexts. A general tendency is that non-heuristic rules tend to have more global contexts than rules in heuristic sections. The reason why even morphology needs global rules is not only the variable size and structure of phrases, but also the fact that syntax - e.g. in the form function words and verbs' combinatorial potential, may play an indirect role in PoS rules, too.

One can conclude that for the annotation of speech data it is is paramount to provide the parser with some kind of delimiter clues concerning clause and phrase structure, if its global rules are to work optimally. Given the pre-existing markup of prosodic breaks in our corpus, these were the obvious choice of delimiter candidates. However, alternative strategies - albeit somewhat more heuristic - are possible even in the absence of such markup. Thus, Bick (1998) introduced the idea of "dishesion markers" based on pauses, stress markings and hesitation interjections (eh, éh), which

were tagged and disambiguated as either <break> or <pause> tags, where the former constitutes a clause or sentence break, while the latter does not, and is allowed inside clauses and even phrases. Dishesion markers can be inserted near prosodic annotation features such as intonation, pauses, interjections etc., but can also be mapped on regular words, using ordinary CG context rules to define syntactic edges, such as conjunctions for clauses, prepositions for pp's and articles and determiners for noun phrases.

For the C-ORAL-Brasil corpus this technique was implemented exploiting the corpus' explicit prosodic markers. The // "major break" was substituted with a semicolon, while the / "soft break" was re-tagged as a comma with two potential readings, <break> and <pause>, where only the former represents a syntactic break, while the latter is allowed inside phrases and between verb and complement. CG-rules were written to distinguish between these two readings, and the comma replaced with a meta-tag for the <pause> cases - making it invisible to ordinary CG disambiguation rules. Contextually disambiguating the function of prosodic breaks allowed us to strike a balance between simply ignoring such markup on the one hand, and syntactic over-segmentation on the other. If the original parser rules were to work optimally, they needed a comma marker that was as close to a standard written language comma as possible.

The following are clear-text versions of some of the CG rules used for this disambiguation task:

- a prosodic /-marker is treated as <break> if it occurs before the first word of an np, or before a pronoun in the nominative, followed by a finite verb to the right (i.e. clause-initially)
- between a noun or a nominative pronoun to the left, and a finite verb to the right, a prosodic /-marker is treated as <pause> (subject - verb case)
- prosodic /-markers between a noun and another np are treated as <break> (appositions)
- prosodic /-markers between potential np-parts are treated as <pause> if there is gender-number agreement between the np part candidates (e.g. DET-ADJ-N)
- between a noun and an adjective agreeing in gender and number, a prosodic /-marker is treated as <pause> (i.e. N ADJ)
- between a transitive verb and a left np edge, a prosodic /-marker is treated a <pause>
- between an auxiliary and its main verb, a /-marker is treated as <pause>
- if a single word is surrounded by prosodic /-markers, these are treated as <pause>
- if a prosodic /-marker is preceded by a conjunction or relative, it is treated as <pause>
- if a prosodic /-marker is preceded by a preposition, it is treated as <pause> (i.e. pp-internal)
- between an intesifier and an attribute, '/' is treated as <pause> (i.e. adjective phrase-internal)
- if a prosodic /-marker is followed by certain, typically postnominal prepositions (de, em, com, sem) in certain contexts, it is treated as <pause>

Of course, since this rule section had to be run *before* the parser's own rules (which it was supposed to help), linguistic context conditions had to be worded carefully and not too explicitly, taking into account the high morphological and PoS ambiguity of raw text input. Within PALAVRAS ordinary grammar, <break> tags will function like commas, in both set definitions, BARRIER conditions and ordinary context conditions.

6. Evaluation

In order to evaluate the modified parser on our data, one transcription file (bfamdl15) was chosen at random, automatically analyzed and hand-corrected. We then used the Constraint Grammar evaluation tool eval_cg to compare the raw analysis file with the revised version. In an ordinary CG setup, meta-markup and punctuation would align 100%, but in our case, matters were complicated by the fact that "commas" had been disambiguated as either break or pause, and in the latter case replaced with a meta-tag. On the one hand, this caused alignment problems for the evaluator, on the

other hand, differences had to be identified and counted as recall errors. Other mismatches, caused by faulty splitting or non-splitting of ambiguous MWE's, were also counted as recall errors, e.g in the case of *"primeiro=que"* (conjunction vs. adjective/numeral + relative). Including "punctuation" tokens, the file contained 1895 word tokens.

	Recall	**Precision**	**F-Score**
Syntactic function	95.3	94.9	95
PoS / Word class	98.5	98.7	98.6
Morphology	98.4	98.6	98.5
Base form	98.6	99.4	99

It can be seen from these figures that the easiest task was lemmatization (base forms), while syntactic function was the most difficult. The difference between recall and precision for syntax is a measure of remaining ambiguous tags. For word class and morphology, only one reading was allowed, so the precision-recall differences are entirely due to differences in matching differences between break markers (commas).

In order to judge the effectiveness of using prosodic break markers as punctuation, we also compared the standard run (with pause/break disambiguation) with a no-break run (/-marks ignored), a no-sentence run (both /, + and // ignored), and an all-break run (all /-marks turned into commas, *without* disambiguation). Since the gold file did have disambiguated commas, the evaluator was run in match-only mode, comparing tags only for matching tokens. Therefore, figures in the table below can only be compared with each other, and not with the original test run[9].

	no-sentence	**no-break**	**all-break**	**pause / break**
Syntactic function	86.2 (R: 86.5, P: 86.1)	90.7 (R: 91.0, P: 90.6)	93.7 (R: 93.3, P: 93.6)	**95.0** (R:95.3, P: 94.8)
PoS / Word class	98.3	98.8	99.3	**99.4**
Morphology	98.1	98.6	**99**	98.7
Base form	99	99.1	**99.4**	**99.4**

Clearly, exploiting prosodic break markers did improve performance at all levels. However, the effect was much more marked for syntax than for part of speech, lemmatization and morphology, reflecting the wider contextual scope of syntactic tags and the ensuing greater need for precise and correct segmentation[10]. Interestingly, while syntactic performance can be further increased by pause/break disambiguation, this is not obvious for the more local tag categories. Thus, for inflexion tags (morphology), all-break performance was *higher* than for the pause/break run, and only for part of speech a slight improvement was observed.

7. Conclusion

While the C-ORAL Brasil annotation project has shown that a standard written-language parser (PALAVRAS) can be used to assign morphosyntactic tags to transcribed speech data, it also demonstrated that for optimal performance, certain adaptations should be made to both the system and the data, comprising some orthographical normalization and lexicon extensions, as well as

[9] Also, this experiment was done at a later stage, where some improvements in the general grammar had been made, making a direct comparison impossible.

[10] It can be concluded that the positive effect of such segmentation on uniqueness principle rules and NOT-rules (which profit from more fine-grained segmentation) outweighs the potential comma-blocking of *positive* rules looking for long-distance syntactic relations.

syntactic segmentation. The latter proved especially important for syntax, and was achieved by exploiting prosodic break markers as "punctuation", enhaced by a rule-based distinctions between pause and break functions. Under optimal conditions, the modified parsing system achieved correctness rates (F-scores) of 98.6% for part of speech, 95% for syntactic function and 99% for lemmatization.

The implemented annotation scheme does preserve the original prosodic-transcriptional information, including speech flow, retractions, overlaps, turntaking etc., encoded as meta-tagging alongside the morphosyntactic tags, but it will be a future task to figure out an integrated search formalism (GUI interface) that would allow the user to work with these two different levels of annotation at the same time, rather than separately. Long-term, higher levels of grammatical annotaion could also be added and made searchable, such as dependency trees, semantic classes, case roles or anaphoric relations, all in principle available for written-language PALAVRAS parses.

References:

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

Bick, Eckhard. 1998. Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese, in: Proceedings of the 17th Scandinavian Conference of Linguistics (Odense 1998)

Bick, Eckhard & Marcelo Módolo. 2005. Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In: Claus Pusch & Johannes Kabatek & Wolfgang Raible (eds.) *Romance Corpus Linguistics II: Corpora and Historical Linguistics (Proceedings of the 2nd Freiburg Workshop on Romance Corpus stics, Sept. 2003)*. pp. 271-280. Tübingen: Gunther Narr Verlag.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language. HLT '91, Morristown, NJ, USA: Association for Computational Linguistics, pp.112–116

Castilho, Ataliba de (ed.), 1993. Gramática do Português Falado, vol.3, Campinas: Editora da Unicamp.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto. 1995. Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter.

Maamouri, Mohamed et al. 2010. From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In: Proceedings of LREC 2010, Valletta, Malta, May 2010.

Moreno, A. & J.M. Guirão. 2003. "Tagging a spontaneous speech corpus of Spanish". In: Proceedings of the International Conference on Recent Advances in Natural Language Processing.). Borovets, Bulgaria, 2003. p. 292-296.

Müürisep, Kaili and Uibo, Heli (2006). "Shallow Parsing of Spoken Estonian Using Constraint Grammar". In: P.J.Henriksen & P.R.Skadhauge, Proceedings of NODALIDA-2005 special session on treebanking. Copenhagen Studies in Language #33/2006

Moneglia, M., A. Panunzi, E. Picchi, 2004, Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus : C-Oral-Rom Italian. In M.T. Lino et al. (eds.), Proceedings of the 4th LREC Conference, vol. 2, ELRA, Paris, pp. 563-566.

Raso, Tommaso & Heliana Mello. 2010. The C-ORAL BRASIL corpus. In: Massimo Moneglia & Alessandro Panunzi (eds): Bootstrapping Infromation from Corpora in a Cross-Linguistic Perspective. Universitá degli studi di Firenze, Biblioteca Digitale.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing 1994. pp. 44-49.

Appendix: Tag definitions

1. Verbs

V Verb

person/number
1S first person singular
2S second person singular
3S third person singular
1P first person plural
2P second person plural
3P third person plural
0/1/2/3S zero-morpheme infinitive

mood
IND indicative
SUBJ subjunctive
COND conditional
IMP imperative

tense
PR present
IMPF past (imperfeito)
FUT future (simple)
PS past (perfeito simples)

non-finite forms
INF infinitive (+ number/gender)
GER gerund
PCP participle (+ number/gender)

2. Nominals and prenominals:

N Noun
PROP Proper noun
ADJ Adjective
DET Determiner
NUM Numeral

gender
M masculine
F feminine
M/F invariable/underspecified

number
S singular
P plural
S/P invariable/underspecified

secondary nominal classes
<KOMP> comparative
<SUP> superlative
<NUM-ord> ordinal
<card> cardinal

PERS Personal pronoun

person/number
1S first person singular
2S second person singular
3S third person singular
1P first person plural
2P second person plural
3P third person plural

gender
M masculine
F feminine
M/F invariable/underspecified

case
NOM nominative
ACC accusative
DAT dative
PIV prepositive
NOM/PIV underspecified
ACC/DAT underspecified

INDP Independent (non-inflecting)

Secondary pronoun classes

<arti> DET indefinite article
<artd> DET definite article
<dem> DET, INDP demonstrative
<poss...> DET possessive
<quant> DET quantifier
<rel> DET, INDP relative
<interr> DET, INDP interrogative
<refl> PERS reflexive personal pronoun
<si> DET reflexive possessive

3. Non-inflecting word classes

ADV Adverb
<rel> relative
<interr> interrogative
<ks> similar to subordinatin conjunction
<kc> similar to coordinating conjunction
<foc> focus marker

PRP Preposition
KS Subordinating onjunction
KC Coordinating onjunction
IN Interjection

4. Syntactic tags

@SUBJ> @<SUBJ subject
@ACC> @<ACC accusative (direct) object
@DAT> @<DAT dative object (only pronominal)
@PIV> @<PIV prepositional (indirect) object
@ADVS> / @SA> @<ADVS / @<SA adverbial predicative (place, time, duration, quantity), equivalent to nominal @SC
@ADVO> / @OA> @<ADVO / @<OA adverbial object predicative, equivalent to nominal @OC
@SC> @<SC subject predicative

@OC> @<OC object predicative
@ADVL> @<ADVL adverbial
@PASS> @<PASS agent of passive
@ADVL 'free' adverbial phrase (in non-sentence expression)
@NPHR 'free' noun phrase (in non-sentence expression without verbs)
@VOK 'vocative' (e.g. 'free' addressing proper noun in direct speech)
@>N prenominal adject
@N< postnominal adject
@N<PRED postnominal (in-group predicative)
@APP identifying apposition
@>A prepositioned adverbial adject
@A< postpositioned adverbial adject
@PRED> 'forward' free predicative
@<PRED `backward' free predicative
@P< argument of preposition
@S< sentence-related “apposition”
@FAUX finite auxiliary (cp. @#ICL-AUX<)
@FMV finite main verb
@IAUX infinite auxiliary (cp. @#ICL-AUX<)
@IMV infinite main verb
@PRT-AUX< verb chain particle (preposition or "que" after auxiliary)
@CO coordinating conjunction
@SUB subordinating conjunction
@KOMP< argument of comparative (e.g. "do que" referring to *melhor*)
@COM direct comparator without comparative
@PRD role predicator (e.g. "work *as*")
@FOC> @<FOC focus marker ("gosta *é* de peixe.")
@TOP topic constituent ("*Esse negócio*, não gosto dele.")
@#FS- finite subclause (combines with clausal role and intraclausal word tag, e.g.@#FS-<ACC @SUB for "não acredito *que* seja verdade")
@#ICL- infinite subclause (combines with clausal role and intraclausal word tag, e.g. @#ICL-SUBJ> @IMV in "*consertar* um relógio não é fácil")
@#ICL-AUX< argument verb in verb chain, refers to preceding auxiliary
@#AS- averbal (i.e. verb-less) subclause (combines with clausal role and intraclausal word tag, e.g. @#AS-<ADVL @ADVL> in "ajudou *onde* possível")
@AS< argument of complementiser in averbal subclause

5. Some secondary tags

<*> upper case
<*1> left quote
<*2> right quote
<sam-> first part of contraction
<-sam> second part of contraction
<parkc-1> first part in *ou...ou*
<parkc-2> secnond part in *ou...ou*
<pp> complex adverb (prepositional phrase)
<hyfen> hyphenated word
<newlex> from add-on lexicon

PALAVRAS also uses some 200 semantic prototype tags for nouns, not listed here, as well as valency tags for verbs, nouns and adjectives.