Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora

Diana Santos, Cristina Mota

Linguateca, Linguateca
SINTEF ICT, FCCN
Diana.Santos@sintef.no, cmota@ist.utl.pt

Abstract

In this paper, we present a system to aid human annotation of semantic information in the scope of the project AC/DC, called `corte-e-costura`. This system leverages on the human annotation effort, by providing the annotator with a simple system that applies rules incrementally. Our goal was twofold: first, to develop an easy-to-use system that required a minimum of learning from the part of the linguist; second, one that provided a straightforward way of checking the results obtained, in order to immediately evaluate the results of the rules devised. After explaining the motivation for its development from scratch, we present the current status of the AC/DC project and provide a quantitative description of its material in what concerns semantic annotation. We then present the `corte-e-costura` system in detail, providing the result of our first experiments with the semantic fields of colour and clothing. We end the paper with some discussion of future work as well as of the experience gained.

1. Overview

Human annotation is considered fundamental for most natural language processing tasks, but is also recognized as tedious and time consuming. Our goal with the work described in the present paper was to investigate how to leverage on the human annotation effort, by providing the annotator with an intuitive and minimal system that applies rules incrementally.

Our goal was twofold: first, to develop an easy-to-use system that required a minimum of learning from the part of the linguist (annotator); second, one that provided a straightforward way of checking the results obtained, in order to immediately evaluate the results of the rules devised.

We were **not** interested in developing yet another large and powerful environment for the sake of elegance or design. Rather, we wanted to experiment with the human-machine workload distribution in corpus annotation. Also, by putting the system in production at an early stage we were interested as well in eliciting further requirements, in close connection with the user(s). And, obviously, we wanted to improve the corpus and tool resources that Linguateca has been providing to the Portuguese language processing community for more than ten years (Santos, 2009; Santos, 2010) in the first place.

The system, named `corte-e-costura`[1] was created in the scope of the AC/DC project (Santos and Bick, 2000; Costa et al., 2009), which allows the free querying on the Web of ca. 25 different parsed corpora containing ca. 250 million words, which have also some semantic information.

The particular task for which `corte-e-costura` was originally created was aiding the human annotation of the corpora for specific semantic fields.

[1]This is the Portuguese name for the traditional activity of changing and adjusting clothes, fixing and improving by modest means, instead of creating brand-new clothes.

1.1. Motivation

Why yet another annotation system? Well, the first answer is that there are not so many annotation systems that are specifically tailored for Portuguese, and it has been our experience for a long time now that off-the-shelf programs always require a set of "corte e costura" procedures to work with our language anyway (Santos, 1999; Mota, 2004).

The second reason is that we were not looking for annotation systems for free text at all. We wanted to add annotation to already heavily processed text. In fact, the combination of the many different attributes is probably unique to our website. And then we did not want fully automatic systems, but flexible systems easy to revise, for people who had been trained in the exact combination of material we wanted to annotate further.

It was therefore a well-motivated choice to start by a minimum program and see how it could help linguists to specify simple rules both for automatic annotation, and for the revision and correction of the results of their own rules.

As for linguistic motivation, we were intrigued by Halliday's proposal (Halliday, 1991) that quantitative data on language was often an issue of 1:9. Halliday argues that frequency properties are as important as the values themselves in a language system, and that they are part and parcel of the same phenomenon. He claims that there are two kinds of frequency patterns in natural language, corresponding to the traditional concepts of opposition and markedness, corresponding to equal probability, and to 10%-90% occurrence. His proposal is in addition reinforced by the argument that, since natural languages have to be learned, it would be hard to learn a different specific probability for each different grammatical or lexical phenomenon.

If this is the case – and our own annotation experience did not go counter it – one would expect that if the annotation was well organized, for each level of rule annotation, only 10% of the cases would have to be corrected, and this until no more refinements could be done.

1.2. Related work

We are perfectly aware that other formalisms and environments might help us with this issue: the constraint grammar formalism (Karlsson et al., 1995) in which PALAVRAS is built, or NooJ (Silberztein, 2003), or even programs that have been created as useful extensions for the IMS CWB, as MP (Schulze, 1996).

The down size of all these choices was to require the linguist to install, and/or get familiar with, any of these systems, when we were not yet sure of how useful would be the rule writing.

In Santos and Ranchhod (1999) we made also the related point that corpus processing has many uses and purposes, and that different corpus processing systems may be developed for different corpus-related activities.

2. The AC/DC project

The AC/DC[2] project was started in 1998-99, and had as its main purpose to make a large number of corpus resources for Portuguese available on the Web, with a unified and simple interface. This interface allowed people to interact with corpora without requiring physical access to institutions or software installation – at that time, there was no such thing for Portuguese.

Later on, we also considered as one of Linguateca's tasks to create new resources, such as a large newspaper text corpus, CETEMPúblico (Santos and Rocha, 2001), and more complex evaluation resources with further information associated, as the CHAVE collection for information retrieval (Santos and Rocha, 2005), and HAREM's golden collections (Rocha and Santos, 2007); all of these resources were later included in AC/DC.

The AC/DC project also widened in other directions, constituting what we came to call the "AC/DC cluster", namely, a set of other related projects that shared the processing core of AC/DC – and in some cases, even the material. So we can consider as members of this cluster also:

- the Floresta Sintáctica[3] treebank – the first treebank for Portuguese, a cooperation with Eckhard Bick and his VISL project at the University of Southern Denmark (Afonso et al., 2002; Freitas et al., 2008);
- the COMPARA[4] corpus – a large manually revised Portuguese-English fiction parallel corpus, a project led by Diana Santos and Ana Frankenberg-Garcia, which finished in December 2008 (Frankenberg-Garcia and Santos, 2003);
- CorTrad[5] – a parallel (multi-version and multigenre) corpus, in cooperation with Stella Tagnin and Elisa Duarte Teixeira and the COMET project, and with NILC (Tagnin et al., 2009).

[2]The name stands for *Acesso a Corpos, Disponibilização de Corpos* (roughly: access to corpora, making corpora available), and is meant to indicate that it should both benefit users – granting them access; and corpus owners: helping them to make their corpora widely available. See `www.linguateca.pt/ACDC/`.

[3]See `http://www.linguateca.pt/Floresta/`.

[4]See `http://www.linguateca.pt/COMPARA/`.

[5]See `http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html`.

2.1. Description of the material

It is always difficult to provide a good overview of what is available for browsing, given that several corpora share material among them, and some have more markup than others. In Table 1, the AC/DC material is roughly quantified under the genre parameter, as of March 2010. No repeated material was included, i.e., overlapping parts were removed from the computations present in the next tables.

Genre	Size in words
Narrative fiction	17,216,517
General newspaper	246,112,528
Specialized newspaper	6,354,317
Informative, technical	4,488,973
Oral	500,741
Other or not classified	1,903,662

Table 1: Genre distribution in the AC/DC cluster

Table 2 presents the material in terms of language variety, only for those corpora where we have this information: for example for the CoNE corpus, a corpus of spam email messages where we have not classified the origin, and therefore the variety, we do not have this information.

Language variety	Size in words
Africa	27,618
Brazil	58,893,616
Portugal	215,321,065
Unmarked/unknown	688,394

Table 2: Variety distribution in the AC/DC cluster

The most obvious presentation of the AC/DC contents is a list of the sizes and short descriptions of their corpora, as displayed in the website and in Table 3.

2.2. Semantic annotation

After having the corpora syntactically annotated, a next step was to do semantics. Because there was no off-the-shelf program that we could have employed, we proceeded in a stepwise way, trying out specific semantic domains first.

We started by the colour domain in Linguateca for several reasons: colour words are common in natural languages, at least in Portuguese, and there is a huge literature on colour, both in English or in languages in general, as well as in Portuguese (see Inácio et al. (2008) for references). Furthermore, we were also interested in classification of images by natural language means. Arguments for the existence of syntactic restrictions on the use of colour adjectives have also been advanced (Ellis, 1993), therefore making colour a category which is also reflected in syntax.

The first corpus to be fully annotated with colour was COMPARA, as described in detail in Santos et al. (2008), Silva et al. (2008a), and Silva et al. (2008b). This was due both to our contrastive interests and to the fact that it had

Corpus	Units	Words	Sentences	Var.	Short description
AmostRA	127.832	98.786	4.965	B	Pos-tagged sample of NILC Corpus
ANCIB	1.690.376	1.258.764	80.992	B	(Moderated) Brazilian discussion list on librarianship
Avante!	7.766.418	6.501.257	204.414	P	Portuguese party-political newspaper Avante!, 1997-2002
CD HAREM	222.407	147.077	8.185	all	Golden collection of the First HAREM
CETEMPúblico	232.543.379	189.575.095	7.665.410	P	Major Portuguese daily newspaper, 1991-1998
CHAVE	123.868.725	99.478.954	4.740.448	P B	News from Público and Folha de São Paulo, 1994-1995
ClassLPPE	1.922.601	1.304.282	74.690	P	Portuguese fiction, drama and poetry, 16th and 19th century
CONDIVport	7.088.775	5.577.632	328.214	P B	Sports, fashion and health magazines (50s, 70s and 2000)
CoNE	925.230	685.225	31.561	any	Spam or general e-mail messages
DiaCLAV	7.758.467	6.651.549	232.152	P	Four Portuguese regional daily newspapers
ECI-EBR	917.127	724.015	44.381	B	Corpus Borba-Ramsey of Brazilian Portuguese
ECI-EE	32.034	27.140	839	P	Call for the EU ESPRIT program
ENPCPUB	92.693	72.389	4.371	P B	Translated fiction from English, subset of the ENPC corpus
FrasesPB	23.313	19.162	653	B	Individual sentences in Brazilian Portuguese
FrasesPP	20.049	16.233	594	P	Individual sentences pos-tagged
MuseuPessoa	517.747	375.158	27.288	P B	Transcriptions of oral interviews from Museu da Pessoa
Natura/Minho	2.156.187	1.749.083	68.910	P	Unedited version of regional Portuguese newspaper
Natura/Público	7.369.349	6.274.542	225.752	P	Two first paragraphs of each Público article, 1991-1994
NILC	42.608.038	32.342.456	1.963.795	B	Newspaper, commercial letters,textbooks, etc.
Vercial	18.854.273	14.315.992	596.869	P	Portuguese fiction, poetry and drama, 16th to 20th century
Total	457.707.958	368.107.085	16.342.734		

Table 3: AC/DC corpora: Portuguese variety is marked P for Portugal and B for Brazil. Much more information on each corpus is on the website.

undergone several levels of revision, from the texts itselves to the alignment to the syntactic annotation.

We then reused the data (colour terms and classes) and knowledge gathered with COMPARA to annotate all other corpora in the AC/DC cluster, that is, in this case all AC/DC corpora and CorTrad as well.

We started by the CONDIV corpus (Soares da Silva, 2008; Soares da Silva, 2010) not only because it contained very different genres from literary fiction, namely texts on football, on fashion and on health, but also because the way it was built allowed us to compare systematically two varieties of Portuguese, three temporal areas, and three subject domains. A description of colours in the three genres can be found in Table 4. We stress in any case that it is preliminary, because it was not yet 100% humanly revised.

We defined the statistics I_{tok} and I_{typ} as indices that indicate relative colour extension and variety. I_{tok} stands for the ratio of the number of colour tokens to the total number of word tokens in the corpus (times 10,000), while I_{typ} is likewise defined. So, a large number in both indices indicates a lot of colour information being present. Also, for the same I_{tok}, a large I_{typ}, in addition, indicates that the colour field in itself is relevant for the texts. Finally, the ratio of colour types to colour tokens (colour type-token ratio, CTT) indicates variety of shades as opposed to frequency of mention.

For types, we use the lemma values, so cases like *amarelo*, *amarela*, *amarelíssimo*, *amarelos* and *amarelinho* are all subsumed under the AMARELO (adjective) type. On the other hand, the two other forms *amarelo* classified as noun with lemma AMARELO and verb with lemma AMARELAR are counted as different types.

We present in Figure 1 the two indices for all AC/DC corpora. It should be noted that these values have to be treated with care since they were computed for data of very different sizes and are therefore not easily comparable, cf. Baayen (2001) for a discussion of these and similar matters. We are currently investigating different ways of better visualizing and comparing these corpora.[6]

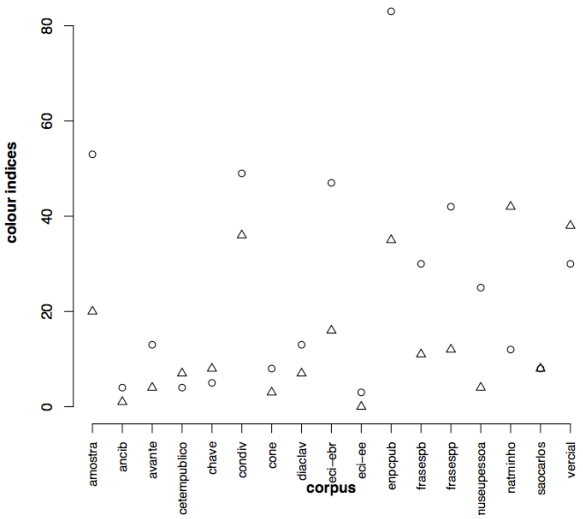

Figure 1: The semantic field of colour in AC/DC: circles correspond to I_{typ}, triangles to I_{tok}.

After colour, it was natural to start as well by marking clothing due to the fashion theme of CONDIV and the fact that there were already clothing terms available from the CONDIV project.[7] Football terms, on the other hand, were

[6]Two obvious strategies are: select an equal sized random sample for every corpus, and/or estimate theoretical values corrected by the corpus size.

[7]Also, there were other sources of clothing information, such as the TexSITE project, `http://pt.texsite.info/` and WordNet.PT, and studies we knew about (Bacelar do Nascimento

Subject	Tokens	Types	Colour tokens	Colour types	I_{tok}	I_{typ}	CTT ratio
Fashion	1,104,347	33,787	13,763	477	124.626	141.179	.0346
Health	1,806,201	46,948	2,276	160	12,601	34,080	.0703
Soccer	2,992,062	60794	4457	120	14,896	19,739	.0269

Table 4: Colour distribution in CONDIV. CTT stands for Colour type/token (ratio).

deemed to be less relevant and conspicuous in all sorts of corpora, so they were not yet dealt with.

Clothing annotation is still at an early stage, cf. Santos et al. (2010). We have, in any case, already defined a somewhat different organization of classes that appear relevant in this context, since different distinctions and properties seem to be at play in Portuguese:

- Clothes are not as widespread grammatically as colours: while we found colours in almost all parts of speech, for clothing the vast majority of items are nominal, with very few verbs indeed: in addition to the prototypical *vestir* and *despir*, and *calçar* and *descalçar*, only a few such as *agasalhar* ('put on warm clothes'), *descobrir* ('take off the hat'), *encasacar-se* ('put too many coats').
- A sort of classifier-like system shows in the pair *calçar* versus *vestir* (to which a neutral *pôr* can be added): in fact, *calçar* implies from below (or to a binary pair of items such as feet, legs or hands), and is used for shoes to trousers but also to mittens, while *vestir* implies from above (or a unique dressing piece) and is used for dresses and everything that is on the upper part of the body but also shirts. Finally, for head garments *pôr* is required.[8]
- As a cultural attribute per excellence, gender issues are intricately mixed with the lexicon of clothing, which naturally reflects differences between women's and men's clothes. There are also special clothes for babies and infants. But there is a marked tendency to have common clothes for the two sexes as well. So we are marking four independent categories as well: `roupa:mulher` (woman), `roupa:homem` (man), `roupa:criança` (child), and `roupa:unissexo` (both genders).
- We have also added an additional layer of classification to the 33 clothing classes currently considered, namely, the six superclasses: `agasalhos` (warm clothes), `calças` (trousers), `casacoscurtos` (short coats), `conjuntos` (more than one piece), `exteriortronco` (pieces used over the shoulders), `vestidoousaia` (clothes including skirt).

As was the case for the colour domain, there is a considerable difficulty in deciding what to annotate in borderline cases, and we proceed by documenting rigorously all options taken, together with the cases for which we take an arbitrary decision. Examples of decisions consistent with the approach taken for the colour are:

- The specification of a `NãoEspecificada` (unspecified) category, including generic references to clothes but not which (just like for colour), including items such as *roupa*, *farrapos*, *vestes*, *fato*, etc.;
- The use of a `Outras` (other) category, to include all cases for which we so far did not think a specific category was warranted. To have a glimpse of the kinds of items here, we have bathing suits; culturally-defined clothes such as *kimono* or *poncho*, and work-clothes such as *fato macaco* (overall), *bata* (school uniform), or *hábito* (monks clothes);
- The existence of expressions that involve clothing but have a metaphorical import, and which are therefore reclassified in a `Original` class. Examples are *de tirar o chapéu* (literally, 'deserves taking off our hats', for something very good) and *apertar o cinto* (literally, 'to squeeze the belt', for describing expense reductions).

Problems that so far are unique to the clothing domain is the much higher vagueness of terms out of context: for example, *casaco* or *fato-macaco* can belong to three or four classes, and the existence of several related domains such as accessories, parts of clothes (such as collars and zips) and materials, which apparently are more natural to classify in the same fell swoop. We have in any case also compiled lists of those lexical items.[9]

2.3. Need for revision

We needed a system to improve the revision process of the annotated corpora of the AC/DC cluster projects. All these projects share a similar workflow in that the corpora are first parsed by PALAVRAS (Bick, 2000) (or a corresponding English system, if bilingual corpora), then they are transferred to the AC/DC format (Santos, 2008) adding meta-information as well, and then they are automatically annotated with information pertaining to several (semantic) domains.

One important feature is that these corpora may be improved (or the annotation may be changed) due to for example correction of the input texts or improvement of the PALAVRAS version or lexicon. So, we did not want to fix their contents and let people start to improve them manually. This would bring severe problems of integration of

and Carvalho, 1996).

[8]This is a system that as native speakers we observe that is currently undergoing change: For example, while for whatever you use *vestir* you can use *despir* for the opposite action, the verb *descalçar* is hardly in use except for shoeware, and there is marked tendency in Portugal to use mainly *tirar* ('remove') for taking off clothes, shoes, etc.

[9]A similar case but less frequent for colour was the case of vegetable materials, such as in *madeira de castanho* (a kind of wood) and the concept of *green* as opposed to *ripe*.

the annotated and revised material with new versions of programs applied in earlier stages of the corpus annotation chain.

We were in addition aware of the time and labour involved in making a complete corpus human revised, since we had followed this route in COMPARA and in Floresta, and we wanted to minimize human effort. In order to make the most of the previous experience of the annotators, the rules should be established by the human annotators in a format as similar as possible to the one used to query the corpus – the IMS-CWB format (Evert, 2009). These rules should modify, delete or add new attributes to the previous annotation, but could be applied automatically as many times as required.

3. The corte-e-costura program

Our goal was therefore to create mechanisms and procedures for semi-automatic semantic annotation that would produce good results in a short period of time, and which could be repeatedly applied whenever changes in corpora would take place.

Currently all semantic information is encoded in two attributes only: `sema`, and `grupo`, which encode a wide range of alternatives, as shown in Santos et al. (2008). However, the program is flexible enough to be applied so that it modifies any attribute or a group of attributes that are encoded in the AC/DC format, and this has actually been extensively used in the correction rules.

On the other hand, at the time of this writing, we have to acknowledge that the program is very slow, and does not yet allow some of the more expressive CWB constructs such as `within s` (restriction to a span described by a structural attribute, in this case `s` for sentence) or `[pos!="N"]*` (negation, regular expressions over values, and/or over tokens).

The bulk of the work so far has been around filling the attributes `sema` and `grupo`, as well as to identify multiword expressions that should be annotated as a single token, which in the AC/DC format are units delimited by the structural attribute `mwe`.

In Silva and Santos (2010) the types of rules used to annotate the colour field, and their linguistic motivation, were presented in detail. A similar documentation for the clothing domain is under way in Santos et al. (2010).

3.1. Rule description

Each rule is comprised of an antecedent and a consequent, which are both composed of one or more terms. When the rule aims at delimiting new multiword expressions and does not depend on the context, the consequent can be omitted. In that case, the antecedent is merely the multiword expression that one wants to annotate, such as *cor de rosa*.

Rules to modify attributes of existing units in the corpus always feature an antecedent and a consequent. In the AC/DC corpora, following the CWB scheme, there are two types of attributes: positional and structural. Both sides of the rule can refer to terms of the two types, allowing simultaneous modification of both kinds.

(1)
```
a:[lema="camisa"] b:[lema="salmão" & pos="N"] >> a:[sema="roupa"]
b:[pos="ADJ" & gen="F" & sema="cor"]
```

(2)
```
[pos="N"] a:<mwe pos="N"> b:[lema="cor"] [lema="de"]
[lema="laranja"] </mwe> >>
a:<pos="ADJ"> b:[sema="cor"]
```

In example (1), upon encountering the sequence *camisa salmão* in the corpus the program adds (or replaces, if the attribute already exists) to the first token the indication that it concerns clothing (`sema=roupa`), and changes the attributes `pos`, `gen` and `sema` of the term *salmão* into the interpretation of this term as a colour term used as an adjective.

A similar action is done in (2), now grammatically (by a noun) and not lexically motivated, and dealing with the multiword *cor de laranja*, initially annotated as a (multiword) noun. Both general attributes and one referring to the first word of the expression are changed by this rule.

Concentrating on multiword expressions, these can be delimited either implicitly or explicitly. In the first case, it suffices to initialize the semantic type and the POS tag that are going to be assigned to the multiword expression independent of context, and list all the instances we want to annotate (initially), as in the next example:

(3)
```
# sema=cor pos=ADJ
açúcar queimado
```

If the multiword expression depends on the context, explicit rules similar to (2) can or should be used, also to create a multiword out of a sequence of words.

(4)
```
[pos="N"] a:[word="peito"] b:[word="de"] c:[word="rola"] >>
<mwe sem="cor" pos="ADJ">
a:[sema="cor"] b:  c:  </mwe>
```

3.2. Rule application

For each sentence S in a given corpus C, the set of rules R is applied using this simple algorithm:

1. Copy S to S'
2. For each element E of S do:
   (a) For each rule in R do:
      i. If rule is activated starting in element E, then execute rule consequent modifying S
3. If S equals S', then return S; otherwise repeat rule application to sentence S from 1.

The repetition step mentioned in step 3. occurs only if the rules are being applied in recursive mode, i.e., when one wants to apply the rules until no more text modifications are observed.

Rules are sequentially applied to each sentence left to right, one at a time, in the same order as enumerated in the rule file. A rule is activated if each term of its antecedent matches an element of the sentence. A rule can be activated more than once per sentence, if there are multiple subsequences that match the rule antecedent. Likewise, there may be more than one rule that matches the same sentence subsequence.

Whenever a rule is activated it is immediately executed. This means that a rule can produce the necessary modification to trigger subsequent rules or to prevent subsequent rules to be activated.
The process advances to the next rule, as soon as there is a rule term that is not satisfied.
Execution of a rule means updating the attributes of the elements of the sequence that triggered the rule. In the antecedent of the rule, a multiword expression can be either explicitly represented by all its constituents or by using the wildcard `*` between the structural tags `<mwe>` and `</mwe>`. For instance, the following rule is triggered when a noun term is preceded by any multiword expression, regardless of the constituents it is built of:

```
<mwe> * </mwe> a:[pos="N"] >>
a:[pos="ADJ"]
```

4. Experiments and results

In this section, we describe the application of the program to the AC/DC corpora, what we learned, and the number of differences that are changed and provide a first idea of the changes in each step.

4.1. Experience gathered

In order to optimize performance and rule writing, we conceptually separated the rules among general rules, and corpus specific rules. Later on we understood that for cases of corpora with very different material, rules for particular subjects were warranted. This, in fact, occurred as well for different varieties.
So, variety differences or domain differences had to be catered for by more specific rule sets: for example, in a soccer context, *amarelo* has a specific interpretation (that of a yellow card) that is not primed in a fashion domain. Conversely, *fato* is a hyperonym for clothes in Portuguese from Portugal but means "fact" in Brazilian Portuguese, while *terno* is only clothing in Brazil and not in Portugal.
It is however important to note that it is not always obvious to consider a rule as general or corpus-specific.
Another thing that required tampering with was the conceptualization among four different kinds of rules (the correction rules are straightforward conceptually, and, in fact, they may not even belong to semantic annotation proper).
We had the following cases:

1. positive rules: for cases where a particular context implies that a given lexical item is positively a colour (or clothing, or whatever domain we are annotating)

2. negative rules: for cases where a particular context implies that a given lexical item – that was marked as default – is not of the semantic domain we are analysing

3. specialization rules: for cases where specialized contexts imply a change in the semantic domain, for example assigning subset classifications such as political colour or team colour

4. recursive rules: to be applied on the end, these rules are not only costly but depend on previous applications, so they are mainly used for coordinations, where the fact that something non-standard is coordinated with other members of a given class allows for disambiguation and/or even for coercion, as is the case of *Quando usar o batom vermelho, que vai bem a qualquer hora, pinte as pálpebras com sombra marrom e* **pérola**. (... paint the eyelids with a brown and **pearl** shadow).

Now, it is important to explain that it is up to the rule writer how she conceptualizes the issues. It is sometimes easier to write negative rules than positive ones no matter the frequency. This was the case with *branqueamento* (whitening, also meaning money washing): although most of the cases referred to the illicit activity concerning money, these cases were easier to identify and thus remove with negative rules than the opposite.

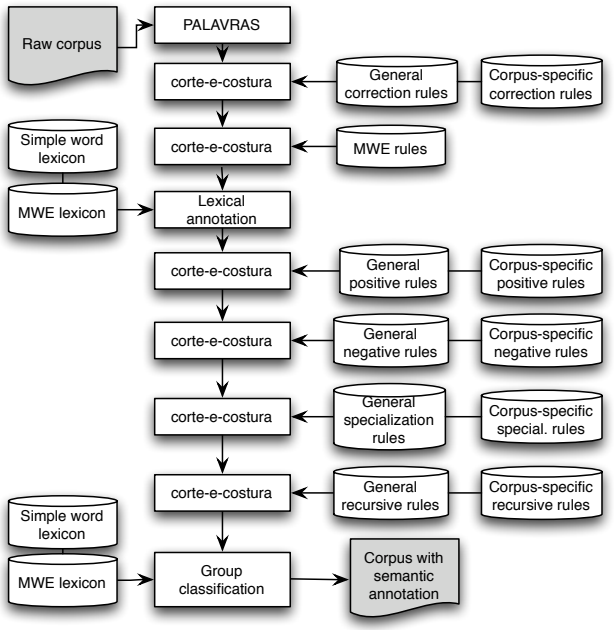

Figure 2: Building blocks for semantic annotation

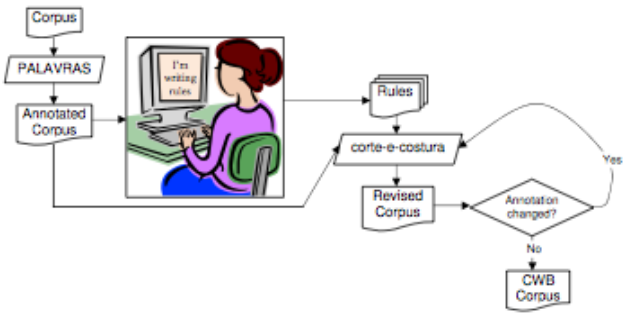

Figure 3: Annotation revision workflow

4.2. Program invocation in the AC/DC

Currently we invoke `corte-e-costura` several times, previously to the creation of a CWB corpus. Although it leads to a considerable slower corpus creation, this allows a clear conceptual organization of the different kinds of rules as well as the possibility to log their application differently.

Figure 2 depicts the invocation as different building blocks, and Figure 3 the overall annotation workflow.

4.3. Number of rules

Although the numbers here are still preliminary since annotation is underway – the reader is redirected to the AC/DC website for news of the annotation and sizes of the resources –, we present here some data on the changes brought from the application of the last version of the several rules. We also note that there are hardly any rules for clothing yet, as Table 5 shows.

Kind of rules	Colour	Clothing
Correction	94	0
Positive	76	3
Negative	47	1
Specialization	109	0
Recursive	10	0

Table 5: Number of rules.

In Table 6 the number of changes in some of the corpora is presented.

5. Concluding remarks

Now that we have learned that this way to proceed does, in fact, allow for an expedite annotation flow, with a minimum learning from the linguist side, we will investigate the two alternative possibilities: (i) improving `corte-e-costura` for this specific task and environment, (ii) rewriting it simply as a frontend to a more powerful system.

We should, in any case, make clear that the process we started is available to every one interested in semantic annotation of Portuguese, who wants to experiment with other semantic domains or other issues.

We welcome and encourage collaboration by the community in the following way: if a researcher is interested in some other subject or domain, provided s/he provides us with lexicons and rules we will annotate the AC/DC corpora with them and make the result available for everyone.

Acknowledgements

Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC, MCTES, UMIC and FCCN.

We thank Rosário Silva for her thorough use of the `corte-e-costura` program and for her dedicated annotation work in the AC/DC project, both in colouring and clothing domains.

Both Rosário Silva and Augusto Soares da Silva are to be thanked as well for the decisions concerning clothing.

6. References

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In Manuel González Rodrigues and Carmen Paz Suarez Araujo, editors, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1698–1703, Paris, 29-31 May. ELRA.

R. Harald Baayen. 2001. *Word frequency distributions*. Kluwer Academic Pusblishers.

Maria Fernanda Bacelar do Nascimento and Anabela Carvalho. 1996. Preto e branco ou branco e preto? (Como se combinam os nomes de cores). In *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística (Lisboa, 2-4 de Outubro de 1995)*, pages 367–380. APL/Colibri.

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Aarhus, Denmark, November.

Luís Costa, Diana Santos, and Paulo Alexandre Rocha. 2009. Estudando o português tal como é usado: o servico AC/DC. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 September.

John M. Ellis. 1993. *Language, Thought and Logic*. Northwestern University Press.

Stefan Evert, 2009. *The CQP Query Language Tutorial*, 4 November.

Ana Frankenberg-Garcia and Diana Santos. 2003. Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors, *Corpora in Translation Education*, pages 71–87. St. Jerome Publishing, Manchester.

Cláudia Freitas, Paulo Rocha, and Eckhard Bick. 2008. Floresta Sintá(c)tica: Bigger, Thicker and Easier. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, pages 216–219, Berlin/Heidelberg, 8-10 September. Springer Verlag.

M.A.K. Halliday. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer and Bengt Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pages 30–43. Longman.

Susana Inácio, Diana Santos, and Rosário Silva. 2008. COMPARAndo cores em português e inglês. In Sónia Frota and Ana Lúcia Santos, editors, *Artigos seleccionados do XXIII Encontro da Associação Portuguesa de Linguística (APL)*, pages 271–286. APL/Colibri.

Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin/New York.

Cristina Mota. 2004. Inflection of the Portuguese DELAS using FST. In Claude Muller, Jean Royaute, and Max Silberztein, editors, *INTEX: Pour la linguistique et le traitement automatique des langues*, number 1 in Archive, Bases, Corpus, pages 35–51. Presses Universitaires de Franche-Comté, France.

Paulo Rocha and Diana Santos. 2007. Disponibilizando a Colecção Dourada (OBRa) do HAREM (ACONTECIMENTO) através do projecto AC/DC (LO-

Corpus	Rules fired	Words changed	Word types changed	Attr. changed	New MWE	MWE changed
AmostRA	30	312	142	570	5	3
ANCIB	93	388	108	641	21	3
Avante!	707	4237	372	7641	40	34
CHAVE	19965	132244	1998	242486	2796	700
CoNE	27	393	105	642	1	0
DiaCLAV	1973	7513	481	13116	444	25
ECI-EBR	150	1969	412	3699	25	3
ENPCPUB	225	377	162	637	6	2
MuseuPessoa	71	751	137	1451	7	1
Natura/Minho	553	1388	224	2093	39	2
Vercial	2921	50246	1553	93831	546	99

Table 6: Number of changes in the corpora: very preliminary data.

CAL—ORGANIZACAO—ABSTRACCAO). In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, pages 307–326. Linguateca, 12 de Novembro.

Diana Santos and Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. In Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, and Gregory Stainhauer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 205–210, 31 May-2 June.

Diana Santos and Elisabete Ranchhod. 1999. Ambientes de processamento de corpora em português: Comparação entre dois sistemas. In Irene Rodrigues and Paulo Quaresma, editors, *Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99)*, pages 257–268.

Diana Santos and Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, 9-11 July.

Diana Santos and Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, pages 821–832. Springer, Berlin/Heidelberg.

Diana Santos, Rosário Silva, and Susana Inácio. 2008. What's in a colour? Studying and contrasting colours with COMPARA. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 28-30 May.

Diana Santos, Rosário Silva, and Augusto Soares da Silva, 2010. *Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português*. http://www.linguateca.pt/acesso/GuardaFatos.pdf.

Diana Santos. 1999. Towards language-specific applications. *Machine Translation*, 14(2):83–112, June.

Diana Santos. 2008. Curso avançado de estudos contrastivos usando o COMPARA como ferramenta. EBraLC, Segunda Escola Brasileira de Linguística Computacional, Universidade Estadual Paulista - UNESP - Campus de São José do Rio Preto.

Diana Santos. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 1(1):25–59, May.

Diana Santos. 2010. Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties. submitted to OSLA.

Maximilian Bruno Schulze, 1996. *MP User's Manual*, 16 April. Institutt für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Max Silberztein, 2003. *NooJ manual*. available at the NooJ web site http://www.nooj4nlp.net.

Rosário Silva and Diana Santos, 2010. *Arco-íris: notas sobre a anotação do campo semântico da cor em português*. http://www.linguateca.pt/acesso/ArcoIris.pdf.

Rosário Silva, Susana Inácio, and Diana Santos. 2008a. Colouring COMPARA: contrastive and monolingual colour studies in English and Portuguese, March.

Rosário Silva, Susana Inácio, and Diana Santos, 2008b. *Documentação da anotação relativa à cor no COMPARA*, 31 December. `http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf`.

Augusto Soares da Silva. 2008. O corpus CONDIV e o estudo da convergência e divergência entre varied ades do português. In Luís Costa, Diana Santos, and Nuno Cardoso, editors, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*, pages 25–28. Linguateca.

Augusto Soares da Silva. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman, editors, *Advances in Cognitive Sociolinguistics*. Mouton de Gruyter.

Stella E. O. Tagnin, Elisa Duarte Teixeira, and Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica*, 4:314–323.