STRUCTURAL LEXICAL HEURISTICS IN THE AUTOMATIC ANALYSIS OF PORTUGUESE

Eckhard Bick

Department of Linguistics, Århus University, Willemoesgade 15 D, DK-8200 Århus N
tel: +45 - 8942 2131, fax: +45 - 86 281397, e-mail: lineb@hum.aau.dk
http://visl.hum.ou.dk/Linguistics.html

Abstract

The paper discusses, on the lexical level, the integration of heuristic solutions into a lexicon based and rule governed system for the automatic analysis of unrestricted Portuguese text. In particular, a morphology based analytic approach to lexical heuristics is presented and evaluated. The tagger involved uses a 50.000 entry base form lexicon as well as prefix-, suffix- and inflexion endings lexica to assign part of speech and other morphological tags to every wordform in the text, with recall rates between 99.6% and 99.7%. Multiple readings are subsequently disambiguated by using grammatical rules formulated in the Constraint Grammar formalism. On the next level of analysis, tags for syntactical form and function alternatives are mapped onto the wordforms and disambiguated in a similar way. In spite of using a highly differentiated tag set, the parser yields

correctness rates - on running unrestricted and unknown text - of over 99% for morphology/PoS and 97-98% for syntax.

After compilation, the system runs at about 200 words/sec on a 200 MHz Pentium based Linux system, when using all levels. Morphological and POS disambiguation alone approach 2000 words/sec. A test site with a variety of applications (parsing, corpus searches, interactive grammar teaching and - experimental - MT has been established at http://visl.hum.ou.dk/Linguistics.html.

1 Background

In corpus linguistics, most systems of automatic analysis can be classified by measuring them against the bipolarity of rule based versus probabilistic approaches. Thus Karlsson (1995) distinguishes between “pure” rule based or probabilistic systems, hybrid systems and compound systems, i.e. rule based systems supplemented with probabilistic modules, or probabilistic systems with rule based “bias” or postprocessing. As a second parameter, lexicon dependency might be added, since both rules based and probabilistic systems differ internally as to how much use they make of extensive lexica, both in terms of lexical coverage and granularity of lexical information.

The *constraint grammar* (CG) formalism (e.g. Karlsson et.al., 1995 or Karlsson, 1994), which I have been using in my own system[1] for the automatic analysis of unrestricted Portuguese text (Bick, 1996 [1] and 1997 [2]), is both rule governed and lexicon based, focussing on disambiguation of multiply assigned lexical and structural readings as the main tool of analysis. Readings are expressed as sets of word based modular tags. Syntactic structure is covered by using function tags and dependency markers (Bick, 1997 [1]), but I will here concentrate on the lexico-morphological level. Before any Constraint Grammar rules can apply, all (morphologically) *possible* readings have to be identified, and I have to this end developed a preprocessor, that identifies wordforms, polylexical units and sentence boundaries, as well as a morphological analiser for Portuguese using an adapted electronic version of a previously published dictionary (Bick, 1993) in combination with affix- and inflection endings lexica supplemented by corresponding alternation rules for word formation (Bick, 1995). In the analyser's output, every word form is followed by as many tag lines as there are potential readings:

(1)
```
"<revista>"
    "revista"  <+n> <CP> <rr> N F S
    "revestir"  <vt> <de^vtp> <de^vrp>  V PR 1/3S SUBJ VFIN
    "revistar"  <vt>  V IMP 2S VFIN
    "revistar"  <vt>  V PR 3S IND VFIN
    "rever"  <vt> <vi>  V PCP F S
```

With a CG-term, such an ambiguous list of readings is called a *cohort*.. In the example, the word form 'revista' has one noun-reading (female singular) and four (!) verb-readings, the latter covering three different base forms, subjuncitve, imperative, indicative present tense and participle readings. Conventionally, PoS and morphological features are regarded as primary tags and coded by capital letters. In addition there can be secondary lexical information about valency and semantical class, marked by <> bracketing.

A *constraint grammar rule* brings the ambiguity problem to the foreground by specifying which reading (out of a cohort of ambiguous readings for a given word) is impossible (and thus to be discarded) or mandatory (and thus to be chosen) in a given sentence-context. For instance, a rule might discard a finite verb reading after a preposition (2a) , or when another - unambiguous - finite verb is already found in the same clause, with no coordinators present (2b).[2]

(2a) REMOVE (VFIN) IF (-1 PRP)

[discard finite verb readings (VFIN) if the first word to the left (-1) is a preposition PRP)]

---

[1] The system was developed in the framework of a Ph.D.-project at Århus University, over a period of three years, drawing on lexicographic research on Portuguese from an earlier Master's Thesis.

[2] Ordinarily, this disambiguation process works on whole cohort lines, i.e. distinguishes between PoS, base form and inflection, but tolerates competing valency options. However, on a higher level of analysis, I have introduced valency and semantical disambiguation, too. This can be very useful for polysemy resolution, like in "rever", where the transitive <vt> - intransitive <vi> distinction has a meaning correlate: 'tornar a ver' [see again] vs. 'transudar' [leak through]. Likewise, "revista" followed by a name <+n> or being read (semantical class <rr>) is more likely to be a newspaper than an inspection (semantical class <CP> for action: +CONTROL, +PERFECTIVE).

(2b) REMOVE (VFIN) IF (*1C VFIN BARRIER CLB OR KC) (NOT *-1 CLB-WORD)

[discard VFIN if there is another unambiguous (C) finite verb (VFIN) anywhere to the right (*1) with no clause-boundary (CLB) og coordinating conjunction (KC) interfering (BARRIER). Discard only if there is no subordinator (CLB-WORD) anywhere to the left (*-1)]

With current software, before an analysis run, all rules are translated into a finite state network by a compiler program, yielding the actual parser. The Portuguese grammar was originally written in the formalism suggested by Pasi Tapanainen's first compiler implementation, but later rewritten to match the notation used in his new CG-2 parser compiler (Tapanainen, 1996).

By applying the rule set several times, the parser renders more and more words in the sentence unambiguous, and in the end, only one reading is left for every word. Since the individual rule can be made very "cautious" by adding more context conditions, and since the last surviving reading will never be discarded, the formalism is very robust. Even imperfect input will yield *some* parse. Unlike probabilistic systems, where "manual interference" as in the introduction of bias on behalf of irregular phenomena often has an adverse side-effect on the overall performance of the parser (due to interference with the ordinary statistical "rules" based on the *regular* "majority" phenomena), Constraint Grammar tolerates and even encourages the incremental "piecemeal " addition of exceptions and context conditions for individual rules (For a comparison of statistical and constraint-based methods see Chanod & Tapanainen, 1994).

2. System Performance

If they can be made to work on free text, rule based systems can achieve very low error rates. While state-of-the-art probabilistic taggers still have error rates of over three percent[3], even for PoS tagging, CG based systems fare somewhat better. For English word class error rates of under 0.3% have been reported at a disambiguation level of 94-97% (Voutilainen, 1992). For my own Portuguese CG system, test runs runs with near 100% disambiguation on fiction and news texts suggest a correctness rate of over 99% for morphology and part of speech, when analysing unknown unrestricted text[4]. For syntax the figures are 98% for classical literary prose (Eça de Queiroz, "O tesouro") and 97% for the more inventive "journalese" of news magazine texts (VEJA, 9.12.1992), as shown in table (3):

(3) System performance on the PoS and syntactic levels:

Text:	*O tesouro* ca. 2500 words		*VEJA 1* ca. 4800 words		*VEJA 2* ca. 3140 words	
Error types:	errors	correct-ness	errors	correct-ness	errors	correct-ness
Part-of-speech errors	16		15		24	
Base-form & flexion errors	1		2		2	
All morphological errors	17	**99.3 %**	17	**99.7 %**	26	**99.2 %**
syntactic: word & phrases	54		118		101	

[3] Compare, for English, (Garside, 1987) on the HMM based CLAWS system, (Francis and Kucera, 1992) on recovering PoS tags from the Brown corpus, Ratnaparkhi's maximum-entropy tagger trained on the Penn Treebank (Marcus et al., 1993) or Brill's stochastic tagger using automated learning (Brill, 1992). For German the Morphy system described in (Lezius et. al., 1996) achieved an accuracy of 95.9%.

[4] The test texts used were not part of the benchmark corpus used to develop the rules, and fresh text chunks were used for every new test. The present grammar, however, still being improved, does incorporate changes made as a result of test run errors.

syntactic: subclauses	10		11		13	
All syntactic errors	64	**97.4 %**	129	**97.3 %**	114	**96.4 %**
"local" syntactic errors due to PoS/morphological errors	- 27		- 23		- 28	
Purely syntactic errors	37	**98.5 %**	106	**97.8 %**	86	**97.3 %**

3. Lexico-morphological heuristics

Yet even in a rule based CG system, heuristics can be quite useful (for English, see Karlsson et. al., 1995). Thus rules are usually grouped according to their "safety", i.e. their statistical tendency to make errors. Less safe rules can be added as a heuristic level on top of a kernel of safe rules, and will be applied *after* these. Also, statistical inspired "rarity tags" (<Rare>) can be added to certain less probable readings in the lexicon, and then referred to by contextual disambiguation rules. A third field for the application of heuristics is on the *analyser level*, i.e. concerns the (lexico-morphological) *input* of the disambiguation rule system. It is this third type of heuristics I am concerned with here.

Since the higher levels of the parsing system (for example, PoS and syntax) are technically rule based disambiguators, they need *some* reading for every word to work on, which is why even word forms that are "unanalysable" for the system (i.e. word forms that can not be reduced to a root found in the analyser's lexicon) need to be given one or more heuristic readings with regard to word class and inflexion morphology. The vast majority of such cases is accounted for by unknown proper nouns (1-2% of all words, depending on text type), while all other lexico-derivational analysis failures together total a frequency of around 0.4%.

3.1. Proper noun heuristics

For the proper noun class, the obvious heuristics is, of course, treating capitalized words as names. Since the tagger looks at one word at a time, analyses it, and then writes all possible readings to the output file, it can only look "backwards" (by storing information about the preceding word's analysis)[5]. Here four[6] cases can be distinguished, the probability for the word being a proper noun being highest in the first case, and lowest in the last:

- 1. A capitalized word in running text, preceded by a another name (heuristic or not), certain classes of pre-name nouns (<title>, e.g. 'senhor', <+n>, e.g. 'restaurante', 'rua', '-ista'-words and others) or the preposition 'de' after another name
- 2. A capitalized word in running text, preceded by some ordinary lower case word
- 3. A capitalized word in running text, preceded only by other capitalized words (The headline case)
- 4. A sentence initial capitalized word[7]

Another distinction made by the tagger is based upon whether or not the word in question can also be given some other (non-name) analysis, and upon how complex this analysis would be, in terms of derivational depth. The name reading is most probable if no known root can be found, and least probable where an alternative analysis can be found without any derivation. Readings where the word's root part is short[8] in comparison to the substring consisting of its derivational morphemes and flexion endings, are also regarded as less probable.

The following table shows in which cases the tagger will choose a (derived) lexical analysis, a (heuristic) proper noun analysis, or both:

---

[5] Even this minimal context sensitiveness is worth mentioning - TWOL-analysers, for instance, never look back at the preceding word.

[6] In an earlier version, cases 1 and 2 were fused, resulting in a somewhat stronger "name bias": because ordinary lower case words would count as pre-name words, too, most upper case words in mid-sentence would get <HEUR> PROP as one of their tags.

[7] The tagger assumes "sentence initiality", if the last "word" is either a question mark, exclamation mark or a full stop not integrated into an abbreviation or ordinal numeral.

[8] To avoid overgeneration, a number of very short lexemes, like the names of letters (tê, zê), have a <nd> (no derivation) tag in the lexicon. These lexemes are completely prohibited for ordinary derivation, - though some also exist in a special, for-derivation-only, orthographic variant, like letter-names (te, ze) that may combine with each other to form productive "phonetic" abbreviations.

(4) proper noun tagging criteria

Preceding context **Competing analysis**	***sentence-initial***	***after only capitalised words: "headline"***	***after lower case word***	***after name or pre-name noun***
underived, pre-name class 'Senhor'	lexical	lexical	lexical	lexical
underived, not pre-name class 'Concordo'	lexical	lexical	lexical (older version: lexical/PROP)	lexical/PROP
long root, derivational 'Palestr-inha'	lexical	lexical/PROP	lexical/PROP	lexical/PROP
short root, derivational 'Cas-ina'	lexical/PROP	lexical/PROP	lexical/PROP	lexical/PROP
none	PROP	PROP	PROP	PROP

Where both the lexical and the proper noun analysis are chosen, the decision is be made at a later stage of analysis by a special set of CG rules. Quantification (5) on 21.806 words from the Borba-Ramsey corpus, containing 452 (2.1%) of (real or supposed) name chains, yielded an error rate of 2% for the PROP class (positive and negative errors combined, shaded in table 5). This is higher than the parser's usual morphological/PoS error rate of under 1%, but one must take into consideration that all 11 errors occured *heuristically*, mostly with lexically unknown words, of which half were spelled incorrectly. Though they would, of course, only be a problem in sentence initial position, it is worth mentioning, that no tagging errors were found for lexicon-registered (i.e. *non*-neuristic) proper nouns.

(5) Proper noun error rates

correct anaysis: ***chosen tag:***	**Proper noun**	**Other, simpel**	**Other, derived**
PROP	79 (17.5%)	0	0
<HEUR> PROP	362 (80.1%)	2 (0.04%)	0
Other word classes	9 (2.0%)	-	-

3.2 "Unanalysable[9] words": typology and statistics

Though much rarer than names, other types of analysis failures (i.e. word forms that can not be reduced to a root found in the analyzer's lexicon) are more difficult to handle, due to their functional diversity and the lack of a clear morphological marker. Table (6) provides an error typology for a 131.981 word literature and secondary literature corpus (The RNP depository of Brazilian literature), containing 604 words unanalysed in the test run. For comparison, language specified percentages for loan word frequency in the larger, mixed Borba-Ramsey corpus (629.364 words, 2599 analysis failures) are given in parentheses.

[9] In this paper I intend "unanalysable word forms" to mean word forms that cannot - by derivation and/or inflexional analysis - be reduced to a root found in the analyser's lexicon. Of course, only part of these - typing errors and foreign language quotes - are *really* unanalysable, while others might be covered by enlarging the lexicon or enhancing the scientific derivation list.

(6) Language distribution and error type in "unanalysable" words

DOMAIN	NUMBER OF TOKENS	PERCENTAGE	
English	77	12.8	(9.3)
French	78	12.9	(3.7)
Italian	10	1.7	(1.5)
Spanish	28	4.6	(0.6)
German	15	2.5	(0.2)
Latin	24	4.0	(2.7)
orthographic variation (European/accentuation)	125	20.7	*Correctables*
other port. orthographic	74	12.3	*Misspellings*
non-capitalised names and abbreviations	37	6.1	***Encyclopedic lexicon failures***
names and name roots	18	3.0	
abbreviations	19	3.1	
root not found in lexicon	119	19.7	***Core lexicon failures***
found in Aurelio[10]	91	15.1	
not found in Aurelio	28	4.6	
derivation/flexion problem	15	2.5	***Affix lexicon failures***
suffix	8	1.3	
prefix	3	0.5	
flexion ending	2	0.3	
alternation information	2	0.3	
other	2	0.3	
SUM	604	100.0	

Three main groups may be distinguished, comprising of roughly one third of the cases each:

a) orthographical errors (shaded in the table, and partially corrected before heuristics proper by an accent module recognizing regular regional spelling variations)
b) unknown and underivable Portuguese words or abbreviations
c) unknown foreing loan words

Of course, a one-register corpus is not typical with regard to loan word distribution. Ordinarily - as the numbers from the Borba-Ramsey corpus show, English has a larger and French a smaller share in the loan word pool. And while nearly inexistent in the literature corpus, scientific domain words can be quite prominent. Cp. the following percentages from the Borba-Ramsey corpus:

(7)

domain	number	percentage	
medical terms	129	5.0%	(of all analysis failures)
botanical terms	45	1.7%	(of all analysis failures)
pharmaceutical names	102	3.9%	(of all analysis failures)

[10] "Novo Dicionário Aurelio" is the largest monolingual dictionary of Brazilian Portuguese.

The overall frequency of words unanalysable for the system, however, is quite stable: For both the literature and the Borba-Ramsey corpora, as well as for VEJA news magazine texts, the figure is roughly 0.4%.

3.3 Analytical morphological heuristics

Sadly, for optimal performance, the three groups would require different strategies. Foreign words appearing in running Portuguese text are typically nouns or noun phrases, and trying to identify verbal elements only causes trouble. In "real" Portuguese words without spelling errors, structural clues - like flexion endings and suffixes - should be emphasized. These will be meaningful in misspelled Portuguese words, too, but, in addition, specific rules about letter manipulation (doubling of letters, missing letters, letter inversion, missing blanks etc.) and even knowledge about keyboard characteristics might make a difference.

Motivated by a grammatical perspective rather than probabilistics, my approach has been to emphasize groups (a) and (b) and look for *Portuguese* morphological clues in words with unknown stems. Since prefixes have very little bearing on the probability of a word's word class or flexional categories, only the flexion endings and suffix lexica are used. As it also does in ordinary morphological analysis, the tagger tries to identify a word from the right, i.e. backwards, cutting off potential endings or suffixes and checking for the remaining stem in the root lexicon (the main lexicon). Normally, for (multiply) *analysed* words, using Karlsson's law (Karlsson 1992, 1995)[11], the Portuguese analyzer would try to make the root as long as possible, and to use as few derivational layers[12] as possible. For (system-internally) *unanalysable* words, however, I use the opposite strategy: Since I am looking for a hypothetical root, flexion endings and suffixes are all I've got, and I try to make their half of the word (the right hand part) as large as possible.

Working with a minimal root length of 3 letters, and calling my hypothetical root 'xxx', I will start by replacing only the first 3 letters of the word in question by 'xxx' and try for an analysis, then I will replace the first 4 letters by 'xxx', and so on, until - if necessary - the whole word is replaced by 'xxx'.[13] For a word like *ontogeneticamente* the rewriting record will yield the chain below. Here, the full chain is given, with all readings it would encounter on its way. In the real case, however, the tagger - preferring long derivations/endings to short ones - would stop searching at the *xxxticamente* -level, where the first group of readings is found. In fact, the adverbial use of an adjectively suffixed word is much more likely than hitting upon, say, a "root-only" noun whose last 9 letters happen to include both the '-ico' and the '-mente' letter chains by chance.

(8)

ontogeneticamente -> no analysis

*xxx**ogeneticamente***
*xxx**geneticamente***
*xxx**eneticamente***
*xxx**neticamente***
*xxx**eticamente***

[11] Karlsson's law states, that of two morpological analyses of different derivational complexity, the one with *fewer* elements is almost always the correct one.

[12] Karlsson's law can be applied to any string of free (i.e. compounding), derivational or inflexional morphemes, but the frequency of ambiguity types with respect to these three elements will differ from language to language - thus, in Portuguese, compounding is much rarer than in most germanic languages, while Swedish, the language for which Karlsson's law was originally formulated, does have compounding, but not as rich an inflexion morphology.

[13] A similar method of partial morphological recognition and circumstantial categorization might be responsible for a human being's successful inflectional and syntactic treatment of unknown words in a known language; the Portuguese word games "collorido" (president Collor & *colorido* - 'coloured') and "tucanagem" (the party of the *tucanos* & *sacanagem* - 'dirty work'), for instance, will not be understood by a cultural novice in Brazil, even if he is a native speakeropean Portuguese - but he will still be able to identify both as singular, the first as a past participle ('-do') and the secon as an abstract noun ('-agem') of the feminine gender.

*xxx**ticamente*** -> suffix '-ico' (variation '-tico') + adverbial ending '-mente'
"ontogene" <DERS -ico [ATTR]> <deadj> ADV

*xxx**icamente*** -> suffix '-ico' + adverbial ending '-mente'
"ontogene" <DERS -ico [ATTR]> <deadj> ADV

*xxx**camente***

*xxx**amente*** -> adverbial ending '-mente' (variation '-amente')
"ontogenetico" <xxxo> <deadj> ADV

*xxx**mente***

*xxx**ente*** -> "present participle"-suffix '-ente'
"ontogeneticamer" <DERS -ente [PART.PR]> ADJ M/F S
"ontogeneticamer" <DERS -ente [AGENT]> N M/F S
-> causative suffix '-entar'[14] + verbal flexion ending '-e'
"ontogeneticam" <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN

*xxx**nte***

*xxx**te***

*xxx**e*** -> verbal flexion ending '-e'
"ontogeneticamenter" <xxxer> V IMP 2S VFIN
"ontogeneticamenter" <xxxer> V PR 3S IND VFIN
"ontogeneticamentar" <xxxar> V PR 1/3S SUBJ VFIN

xxx -> no derivation or flexion
"ontogeneticamente" <xxx> N F S
"ontogeneticamente" <xxx> N M S

Roots with 'xxx' are present in the core lexicon alongside the "real" roots, including the necessary stem alternations[15] for verbs (here, BbCc for different root-stressed forms and AaiD for endings-stressed forms):

(9)

root	word class	alternation subclass	lexeme ID	target of analysis
xxx	<sf>		54572	feminine noun, typically foreign
xxx	<sm>		54573	masculine noun, typically foreign
xxx-	<v-ar>	BbCc	54576	stem-stressed forms of '-ar'-verbs
xxx-	<v-er>	BbCc	54574	stem-stressed forms of '-er'-verbs
xxx-	<v-ir>	BbCc	54575	stem-stressed forms of '-ir'-verbs
xxxa	<sf>		54577	feminine noun, typically Portuguese
xxxar	<amf>		59547	Portuguese '-ar'-adjective*
xxxar-	<vt>	AaiD	54578	endings-stressed forms of '-ar'-verbs
xxxer	<sm>		54666	masculine noun, typically English*
xxxer-	<vt>	AaiD	54579	endings-stressed forms of '-er'-verbs
xxxia	<sf>		54665	feminine noun, Latin-Portuguese*
xxxir-	<vt>	AaiD	54580	endings-stressed forms of '-ir'-verbs
xxxo	<adj>		54581	ordinary Portuguese adjective
xxxo	<sm>		54582	masculine noun, typically Portuguese

Besides the typical stems ending in '-o', '-a' and '-r', default stems consisting of a plain 'xxx' have been entered to accommodate for foreign nouns with "un-Portuguese" spelling. Like many other languages, Portuguese will force its own gender system even onto foreign loan words, so a

[14] This suffix is regarded as a variant of '-ar', and therefore normalized in the DER-tag: <DERS -ar [CAUSE]>.
[15] Here, BbCc for different root-stressed forms and AaiD for endings-stressed forms, with D, for example, meaning a root to be used with future subjunctive endings.

masculine and a feminine case must be distinguished, for later use in the tagger's disambiguation module.

Since the analyser's heuristics for unknown words prefers readings with endings (or suffixes) to those without, and longer ones to shorter ones, verbal readings (especially those with inflexion morphemes in 'r', 'a' or 'o') have a "natural" advantage over what really should be nouns or adjectives, especially when these appear in their uninflected singular base form. Lexicon-wise, this tendency is countered by adding three of the most commonly ignored nominal cases specifically into the lexicon: (a) English '-er' nouns otherwise only taken as Portuguese infinitives, (b) Latin-Portuguese '-ia' nouns otherwise only read as verbal forms in the imperfeito tense, and (c) '-ar' adjectives otherwise analysed only as infinitives.

Rule-wise, verbal readings alone are not allowed to stop the heuristics-machine, it will procede until it finds a reading with another word class. So, the process is set to ignore *verbal* readings on its way down the chain of hypothetical word forms with ever shorter suffix/endings-parts. Thus, the heuristics-machine will *record* verbal readings, but only stop if a noun, adjective or adverb reading is found in that level's cohort (list of readings). In this context, participles and gerunds - though verbal - are treated as "adjectives" and "adverbs", respectively, because they feature very characteristic endings ('-ado', '-ido', '-ando', '-endo', '-indo').

This raises the possibility of the heuristics-machine progressing from multi-derived analyses (with one or more suffixes) to simple analyses (without suffixes) before it encounters a non-verbal reading. In this case, the application of Karlsson's law does still make sense, and when the heuristics-machine hands its results over to the local disambiguation module, this will select the readings of lowest derivational complexity, weeding out all (read: verbal!) readings containing more (read: verbal!) suffixes than the group selected.

In the misspelled French word *'entaente'*, for example, the verbal reading

(10a) "enta" <DERS -(ent)ar [CAUSE]> V PR 1/3S SUBJ VFIN,

from the 'xxxaente'-level, is removed, leaving only underived verbal readings - from the 'xxxe'-level - along with the desired noun singular reading from the 'xxx'-level.

(10b) entaente ALT xxxaente ALT xxxe ALT xxx
"entaenter" <xxxer> V IMP 2S VFIN
"entaenter" <xxxer> V PR 3S IND VFIN
"entaentar" <xxxar> V PR 1/3S SUBJ VFIN
"entaente" <xxx> N F S
"entaente" <xxx> N M S

Since all disambiguation not related to Karlsson's law is referred to the CG-module, the word class choice between V and N will be contextual (and rule based), as well as the morphological sub-choice of mode (IMP - PR) for the verb, and gender (M - F) for the noun. In the prototypical case of a preceding article, the verb reading is ruled out by

(11a) REMOVE (V) IF (-1 ART)

and the gender choice is then taken by agreement rules such as

(11b) REMOVE (N M) IF (- 1C DET) (NOT -1 M)
REMOVE (N F) IF (- 1C DET) (NOT -1 F)

Consider the following examples of "unanalysable" words from real corpus sentences, where the final output, after morphological contextual disambiguation, is given:

(12a) inventimanhas ALT xxxas (also: one ADJ and three rare V-readings)
"inventimanha" <xxx> N F P 'tricks'
itamaroxia ALT xxxia (also: V IMPF 1/3S IND VFIN)
"itamaroxia" <xxx> N F S 'president Itamar + orthodoxy'

(12b) corruptograma ALT xxxograma (3 other NMS-readings removed by local disambiguation)
"corrupt" <DERS -grama [HV]> N M S 'corruption diagram'
araraquarenses ALT xxxenses (3 other ADJ readings removed by local disambiguation)
"araraquar" <DERS -ense [PATR]> <jh> <jn> ADJ M/F P 'from Araraquara'
falocrática ALT xxxtica (1 other AFS-reading removed by local disambiguation)
"falocrá" <DERS -ico [ATTR]> ADJ F S 'phallocracy, reign of the phallos'
ontogeneticamente ALT xxxticamente
"ontogene" <DERS -ico [ATTR]> <deadj> ADV 'by ontogenesis'

(12c) sra ALT xxx (also: N M S)
"sra" <xxx> N F S '=s.-ra - Mrs.'
dra ALT xxx (also: N M S)
"dra" <xxx> N F S '=d.-ra - Dr.'

(12d) sombrancelhas ALT xxxas (also: one ADJ and three rare V-readings)
"sombrancelha" <xxx> N F P '=sobrancelhas - eye brows'
balangou ALT xxxou (also: N M F and N M S)
"balangar" <vt> <xxxar> V PS 3S IND VFIN '=balançou - balanced'
linfadernite ALT xxxite (3 other NFS-readings removed by local disambiguation)
"linfadern" <DERS -ite [STATE]> N F S '=linfadenite - lymphadenoid inflammation)
alfaltada ALT xxxada (only reading)
"alfaltar" <vt> <xxxar> V PCP F S '=asfaltado - paved'

(12e) cast ALT xxx (also: N F S)
"cast" <*1> <*2> <xxx> N M S 'English: cast'

```
gang ALT xxx                    (also: N M S)
    "gang" <*1> <*2> <xxx> N F S 'English: gang'
tickets ALT xxxs                (also: N F P)
    "ticket" <xxx> N M P 'English: tickets'
hijos ALT xxxos                 (also: ADJ M P)
    "tierra" <xxx> N M P 'Spanish: sons'
```

In (12a) and (12b) the parser assigns correct readings to unknown, but wellformed Portuguese words. Since most ordinary words are already represented in the lexicon, or at least derivable from words registered in the lexicon, unknown words will often come from the realms of word games ('itamaroxia', 'corruptograma'), names ('araraquarense') or science ('falocrática', 'ontogeneticamente'), usually involving productive affixes. Depending on the orthodoxy of the fusion process, these affixes may be recognized (12b), or not (12a). Correctly analysed suffixation greatly eases the burden of disambiguation: in all (12b) cases all members of a cohort have the same word class and morphology, making quick, local disambiguation possible. In (12a), where no suffixes are recognized, cohorts will typically cover several word classes, at least one nominal and one verbal. Still, for Portuguese words, flexion endings and - in uninflected words - the word's last letter will almost guarantee that the correct reading is at least *part* of the cohort.

The parser proceeds much in the same way in (12d), with the lowest ambiguity occurring, where larger morphological chunks (morphemes) are recognised, as with the "inflammation-suffix" *'-ite'* and the past participle ending *'-ado'*, and the highest ambiguity where the analysis has to rely on flexion endings alone ('sombrancelhas' and 'balangou', both with cross-word-class ambiguity). What is special about (12d), is the fact, that all forms are misspellings, with (phonetically?) added ('so**m**brancelhas') or simply mistyped letters, as in 'a**l**faltada' where the typists right and left ring fingers have been confused on the keyboard. Even so, with the help of the surviving morphological clues and contextual disambiguation, the parser is able to assign the right analysis in most cases, especially if the words still look Portuguese. The examples seem to corroborate Constraint Grammar's claim that good morphology is the basis for any reasonable (syntactical) parse.[16]

In (12c), 'dra' and 'sra' are not misspellings, but uncommon variants of the more canonical (and longer) title abbreviations 'd.-ra' (doutora) and 's.-ra' (senhora). There is no rule to describe this particular type of variation, so the word forms are treated as "unknown". With the possible exception of the '-a'-ending, both words don't look very Portuguese, and no structure can be found. Since verbs have the highest and nouns the lowest lexicon coverage[17], and since unknown Portuguese three-letter-verbs are virtually unthinkable, the standard analysis for very short words is N with regard to word class, leaving only gender to disambiguation. Here, a preceding feminine article or a following female name will help the CG rules.

(12e), finally, is the hard case - foreign loan words. English 'cast' and 'gang' do not fit with any Portuguese flexion ending, therefore the default reading N is assigned, gender disambiguation relying on NP-context. In 'tickets' the nominal plural-morpheme is recognized, but the stem - 'ticket' - still lacks a Portuguesish last letter, so again, N is chosen for word class. Spanish loan words, being Romance themselves, fare somewhat better, and 'hijos' (an etymological variant of Portuguese

---

[16] Cp. the following quote from *Constraint Grammar* (Karlsson et. al., 1995, p.37):
"The cornerstone of syntax is morphology, especially the language-particular systems of morphological features. Syntactic rules are generalizations telling (a) how word-forms, conceived as complexes of morphological features, occur in particular word order configuations, and (b) what natural classes, "syntactic functions", can be isolated and inferred in such configurations."

[17] In the English CG-system described in (Karlsson et.al. 1995, p. 296), a similar claim is made: *"Because ENGTWOL [i.e. the morphological analyser] very seldom fails to recognize a verb, a verb reading is not assigned [heuristically] without a compelling reason. Word-final 'ed' is a good clue. ...".*

For Portuguese, I have quantified the problem for a stretch of ca. 200.000 words (cp. table 7), showing that nouns account for 73.08% of unknown words (otherwise: 47.38%), and verbs for ca. 8% (otherwise: 38.5%). The bias against verbs is quite strong: Concluding from the above statistics, a Portuguese word unknown to the PALAVRAS lexical analyser is 9 times more likely to be a noun than a verb (and even if it isn't a noun, it's still three times as likely to be something else rather than a verb).

'filhos') qualifies for both plural nouns and adjectives. Of course, resemblances may be misleading, as in English "profession words" in '-er' ('runner', 'gambler') which mimmick Portuguese infinitives. Since this kind of error is especially common within the very complex verbal paradigms, verbal readings - unlike noun readings (which are also favoured by statistics) - are never allowed to be the only ones, as described above. Thus, there is still a chance that contextual information will do the job in the disambiguation module.

In order to test the parser's performance and to identify the strengths and weaknesses of the heuristics strategy of the parser, I have manually inspected 757 "running" instances[18] of lower case word forms where the parser's disambiguation module received its input from the tagger's heuristics module. The first column shows the word class analysis chosen, and inside the three groups (errors, Portuguese, foreign) the left column gives the number of correct analyses, whereas the right column offers statistics about the mistakes, specifying - and quantifying - what the analysis *should* have been.

(13) Word class distribution and parser performance in "unanalysable" words (VEJA news text)

	A) orthographical errors		B) Portuguese words		C) foreign words[19]		all	
analysis	correct	other	correct	other	correct	other	correct	other
N	119	ADJ 8 ADV 8 VFIN 3 PRON 1 DET 1 PRP 1	212	ADJ 3	226	ADV 11 ADJ 3 PRON 2 PRP 2	557[20]	43
ADJ	25	N 8 GER 2	95	N 7	8	-	128	17
ADV	3	-	5	-	-	-	8	-
VFIN	13	N 4 PCP 1 ADV 1	9	N 4 ADJ 2	-	N 7 ADJ 1	22	20
PCP	10	-	16	-	-	-	26	-
GER	3	-	-	-	-	-	3	-
INF	9	-	4	-	-	N 4	13	4
	182	38	341	16	234	30	757	84
		(17.3%)		(4.5%)		(11.4%)		(10.0%)

[18] The words comprise all "unanalysable" word forms in my corpus, that begin with the letters 'a' and 'b'. Since the relative distribution of foreing loan words and Portuguese words depends on which initial letters one works on ('a', for one, is over-representative of Portuguese words, whereas 'x'. 'w' and 'y' are English-only domains), no conclusions can be drawn about these two groups' relative percentages. Inside the Portuguese group, however, the distribution between real words and misspellings may be assumed to be fairly alphabet-independent. Any way, the sampling technique has no significance for error frequencies or distribution in relation to word class, which was the main objective in this case.

[19] Only individual words and short integrated groups are treated, foreign language sentences or syntactically complex quotations are treated as "corpus fall-out" in this table.

[20] This number contains all elements of English noun chains, i.e. the tag N is accepted for all elements in both *death star* and *dead star*, though the second contains what in an English analysis would be an adjective. However, since the English NP in the Portuguese sentence functions as one entity and no analytic Portuguese grammar rules apply inside the term, it seems fair to assign the N-tag to the whole *and* its parts, in the same way foreign name chains are treated as PROP PROP ..., even if one element happens etymologically to be an adjective, as in *United Nations.*

The table shows that, when using lexical heuristics, the parser performs best - not entirely surprisingly - for wellformed Portuguese words (B). Of 323 nouns and adjectives in group B, only 16 (5%) were misanalysed as false positives or false negatives. The probability for an assigned N-tag being correct is as high as 98.6%, for the underrepresented adverb and non-finite verbal class even 100%. All false positive nominal readings (N and ADJ) are still in the nominal class, a fact that is quite favourable for later syntactic analysis.

Figures are lower for group C, unknown loan words, where the chance of an N-tag being correct is only 92.6%, even when allowing for a name-chain-like N-analysis of English adjectives integrated in noun clusters of the type 'big boss'. Finite verb readings, though rare (due to lacking flexion indicators), are of course all failures, and only the little adjective group was a hit, the few cases being triggered by morphologically "Portuguesish" Spanish or Italian words.

The results in group A (misspellings) resemble distributionally those of group B, with a good performance for classes with clear endings, i.e. non-finite verbs and '-mente'-adverbs, and a bad performance for finite verb forms. For the large nominal groups figures are somewhat lower: 84.4% of N-tags, and only 71.4% of ADJ-tags are correct - though most false positive ADJ-tags are still within the nominal range. The lower figures can be partly explained by the fact that misspelled closed class words (adverbs, pronouns and the like) will get the (default, but wrong) noun reading - a technique that works somewhat better and more naturally for foreign loan words (C), which often are "terms" imported together with the thing or concept they stand for, or names. Also, the percentage of "simplex[21]" words without affixes is much higher among the misspellings in group A than in group B, where all simplex words - being spelled correctly - would have been recognised in the lexicon anyway, due to the good lexicon coverage *before* getting to the heuristics module. Therefore, nouns and adjectives in group A lack the structural information of suffixes that helps the parser in group B: 'xxxo' looks definitely less adjectival than 'xxxístico'. In particular, 'xxxo' invites the N/ADJ-confusion, whereas many suffixes are clearly N or ADJ. Thus, '-ístico' yields a safe adjective reading.

4. Special - "deviant" - word class probabilities for the heuristics module

Is it possible, apart from morphological-structural clues, to use "probabilistics pure" for deciding on word class tags for "unanalysable" words? In order to answer this question, I will - in table (14) - rearrange information from table (13) and compare it to whole text data (in this case, from a 197.029 word stretch of the mixed genre Borba-Ramsey corpus). Here, I will only be concerned with the open word classes, nominal, verbal and '-mente'-adverbial.

(14) Open word class frequency for "unanalysable" words as compared to whole text figures

	whole text	"unanalysable" words							
		orthographical errors		Portuguese words		foreign words		all heuristics	
analyses	%	cases	%	cases	%	cases	%	cases	%
N	47.38	131	63.59	232	63.39	237	95.18	600	73.08
ADJ	12.79	33	16.02	100	27.32	12	4.82	145	17.66
ADV[22]	1.26	3 (+9)	1.46	5	1.37	- (+11)	-	8	0.97

[21] "Simplex" words are here defined as words that can be found in the root lexicon without prior removal of prefixes or suffixes. Of course, the larger the lexicon the higher the likelihood of an (etymologically) affix-bearing word appearing in the lexicon, - and thus not needing "live" derivation from the parser.

[22] Only deadjectival '-mente'-adverbs can meaningfully be guessed at heuristically, and therefore only they should enter into the statistics for word class guessing. Also the base line figure of 1.26% for normal text is for '-mente'-adverbs only, the overall ADV frequency is nearly 12 times as high. Since non-'mente'-adverbs are a closed class in Portuguese, the latter will be absent from the heuristics class of wellformed unknown Portuguese words, but in the foreign loan word group and the orthographical error group they will appear in the false positive section of other word classes (numbers

VFIN	24.96	16	7.77	9	2.46	-	-	25	3.05
PCP	4.96	11	5.34	16	4.37	-	-	27	3.29
GER	2.47	3	1.46	-	-	-	-	3	0.37
INF	6.17	9	4.37	4	1.09	-	-	13	1.58
		206		366		249		821	

Among other things, the table shows that the noun bias in "unanalysable" words is much stronger than in Portuguese text as a whole, the difference being most marked in foreign loan words. The opposite is true of finite verbs which show a strong tendency to be analysable. Finite verbs are virtually absent from the unknown loan word group. For the non-finite verbal classes the distribution pattern is fairly uniform, again with the exception of foreign loan words.

As might be expected, among the "unanalysable" words, orthographical errors and correct Portuguese words show a remarkably similar word class distribution.

A lesson from the above findings might be to opt for noun readings and against finite verb readings in "unanalysable" words, when in doubt, especially where no Portuguese flexion ending or suffix can be found, suggesting foreign material. As a matter of fact, this strategy has since been implemented in the system, in the form of heuristical disambiguation rules, that discard VFIN readings and chose N readings for <MORF-HEUR> words, where lower level (i.e. safe) CG-rules haven't been able to decide the case contextually.

5. Conclusion

It can be shown that lexico-morphological heuristics - at least for a morphology-rich language like Portuguese - can be based on structural clues and the systematic exploitation of derivational and inflectional sublexica. Applied to improve analyser recall on the input level of a Constraint Grammar system, the described technique positively contributed to the overall performance of a lexicon based rule governed tagger/parser. Correctness rates of more than 99% were achieved for the morphological/PoS tagger module, with heuristic error rates running at 2% for proper name heuristics and 4.5% for the heuristical analysis of other unrecognized, but correctly spelled Portuguese word forms. In all, heuristic analysis was needed for 80% of all proper nouns (amounting to ca. 2% of running word forms in news text), but for less than 0.4% of non-name word forms. Finally, word class frequency counts suggest that PoS probabilities for "unanalysable" words in Portuguese texts are quite different from those for the language on the whole.

References

Bick, Eckhard, *Portugisisk - Dansk Ordbog*, Mnemo, Århus, 1993, 1995, 1997

Bick, Eckhard, *The Parsing System "Palavras", Documentation*, unpublished Ph.D. project evaluation, 1995, 1997

Bick, Eckhard, "Automatic Parsing of Portuguese", in *Proceedings of the Second Workshop on Computational Processing of Written Portuguese*, Curitiba, 1996

Bick, Eckhard, "Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk", in: Brøndsted, Tom & Lytje, Inger (eds), *Sprog og Multimedier*, Aalborg, 1997

Bick, Eckhard, "Automatisk analyse af portugisisk skriftsprog", in: Jensen, Per Anker & Jørgensen, Stig. W. & Hørning, Anette (eds), *Danske ph.d.-projekter i datalingvistik, formel lingvistik og sprogteknologi*, pp. 22-20, Kolding, 1997

Brill, Eric, "A Simple Rule-based Part of Speech Tagger", in *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, Italy, 1992

given here in parentheses). In the orthographical error group, both '-mente'-adverbs and closed class adverbs can occur, the first as correct ADV-hits, the other usually as false positive nouns (for instance, 'aimda').

Chanod, Jean-Pierre & Tapanainen, Pasi, "Tagging French - comparing a statistical and a constraint-based method", adapted from: *Statistical and Constraint-based Taggers for French,* Technical report MLTT-016, Rank Xerox Research Centre, Grenoble, 1994

Francis, W.N. & Kucera, F., *Frequency Analysis of English Usage,* Houghton Mifflin, 1982

Garside, Roger & Leech, Geoffrey & Sampson, Geoffrey (eds.), *The Computational Analysis of English. A Corpus-Based Approach,* London, 1987

Karlsson, Fred, "SWETWOL: A Comprehensive Morphological Analyser for Swedish", in *Nordic Journal of Linguistics* 15, 1992, pp. 1-45

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

Karlsson, Fred, "Robust parsing of unconstrained text", pp. 97-121, i: Nellike Oostdijk & Pieter de Haan, *Corpus-based research into language,* Amsterdam, 1994

Lezius, Wolfgang & Rapp, Reinhard & Wettler, Manfred, "A Morphology-System and Part-of-Speech Tagger for German", ", in: Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology,* Berlin, 1996

Marcus, Mitchell, "New trends in natural language processing: Statistical natural language processing", paper presented at the colloquium *Human-Machine Communication by Voice,* organized by Lawrence R. Rabiner, held by the National Academy of Sciences at *The Arnold and Mabel Beckman Center* in Irvine, USA, Feb. 8-9, 1993

Tapanainen, Pasi, "The Constraint Grammar Parser CG-2", University of Helsinki, Department of Linguistics, Publications no. 27, 1996

Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto, *Constraint Grammar of English, A Performance-Oriented Introduction,* Publication No. 21, Department of General Linguistics, University of Helsinki, 1992