“Tagging Speech Data” -

Constraint Grammar Analysis of Spoken Portuguese

Eckhard Bick

e-mail: lineb@hum.aau.dk

web-site: http://visl.hum.ou.dk/

Abstract

The paper discusses the automatic grammatical analysis of spoken language data for Portuguese. A Constraint Grammar based tagger/parser for written Portuguese (Bick 1996 and 1997) was used as a point of departure and run on transcribed portions of a Brazilian urban speech corpus, NURC (“*Norma Lingüística Urbana Culta,* e.g. Castilho et.al., 1989 and 1993). Quantitative evaluation of tagging results showed a stable performance (error rates under 1%) for both written and speech data, while the 2-3% syntactic error rate of the original text parser deteriorated considerably when the same rules were applied to preprocessed speech data (8-9% error rate). However, by introducing additional rules and by disambiguating pauses (in-utterance) and breaks (inter-utterance), error rates could be brought down to 4-5%, suggesting the applicability of the CG appoach to (transcribed) speech data.

1. Introduction

The paper raises the question whether automatic grammatical parsers designed for written language input can be made to handle transcribed spoken language data. A Constraint Grammar based tagger/parser for written Portuguese (Bick 1996 and 1997) was used as a point of departure and run on transcribed portions of a Brazilian urban speech corpus, NURC (“*Norma Lingüística Urbana Culta,* e.g. Castilho et.al., 1989 and 1993). By designing a transcription specific preprocessor as well as by adding new rules and modifying old ones, the original system was to be adapted so as to handle both spoken and written language input.

Constraint Grammar (CG) as both a disambiguation based parsing technique and a notationally flat dependency grammar was introduced by Fred Karlsson (Karlsson et. al. 1995) and has been applied to a variety of languages, including - besides English - Swedish

(Birn 1998), Norwegian (Hagen et. al. 1998) and Portuguese (Bick 1996). CG rules are disambiguation rules and syntactic mapping rules that are applied to morphologically analysed (but still ambiguous) text, differing from probabilistic Hidden Markov Model analysers both in their primarily linguistic nature and their wide scope of context conditions, the context window usually being an entire sentence, not only immediately neighbouring words. A typical (simplified) rule for English would for example discard a direct object reading in favour of a subject reading, if there is no transitive verb to the left, and the next word to the right is a finite verb:

```
REMOVE (@OBJ) (0 @SUBJ) (1 C VFIN) (NOT *-1 <vt>)
```

It is this kind of wide context rule one would expect - when moving from written language to speech data input - to be affected by the speech specific lack of punctuation and clear sentence boundaries.

2. The point of departure

The original Portuguese parser yields correctness rates (defined as recall percentages at near 100% disambiguation) - on unrestricted unknown text - of over 99% for morphology/PoS and 97-98% for syntax (Bick, 1997:2), a performance which makes the system suitable for applications like unsupervised parsing, corpus searches, interactive grammar teaching and - experimental - MT (cf. test site at http://visl.hum.ou.dk/), which have so far all been aimed exclusively at written language. While the morphological/PoS tagger module proved quite robust in test runs on spoken language data (with a success rate of around 99% even without additional rules), syntactic analysis fared somewhat worse, with an initial correctness rate of 91-92% for the - rule-wise - unmodified system.

In order to explain this discrepancy between morphological robustness and syntactic failure, hypotheses like the following can be formulated and subsequently tested by changing the system's preprocessor and rule system accordingly:

- In my parser, rules with **morphological targets** mostly use a **shorter context range** (group structure) than those with syntactic targets, cf. table (rule scope). Thus the proportion of rules without and with unbounded contexts is 10 times as high for rules targeting morphological tags than for syntactic targets, and 70-80% of all syntactic rules

stretch their context all the way to the sentence delimiters – making these rules vulnerable to the speech specific absence or vagueness of such delimiters.

table: rule scope

	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**

- **Incomplete utterances** tend to leave group structure intact more often than clause structure, - at least if one doesn't count repetitional modifications/corrections of prenominal modifiers *(essas esses progressos, esta este caminho, da dos nomes)*, where word class adjacency rules can often override agreement rules.
- Speech data **lacks punctuation** and has **unclear sentence window** borders, which is especially bad for syntactic CG analysis which tends to use many unbounded context restrictions (cp 1).
- Speech data is filled with "**syntactic noise**", repetitions and false starts of one- or two-word chunks, as well as pause and phatic interjections *(ahn, uh, eeh etc.)*.

3. Preprocessor tasks

In order to address these problems and improve syntactic performance, a preprocessor was designed with the specific goal of establishing utterance or sentence boundary candidates and removing syntactic noise.

1. Orthography and layout normalisation (character set, line numbers)

2. Repetitions and false starts (automatically commented out by $-signs)

mas é vo/ voluntária né? -->mas é <$vo/> voluntária né?

então então vem tudo aquilo de cambulhada e im/ e im/ im::POSto sobre nós --> então <$então> vem tudo aquilo de cambulhada e <$im/> <$e> <$im/> <stress> imposto sobre nós

com o com o nome da pesso/ da do do escritor lá
--> com o <$com> <$o> nome da <$pesso/> <$da> do <$do> escritor lá

3. Phonetics

* Vowel length markers are removed, e.g. *u::ma pessoa --> uma pessoa*
* In-word stress marking is commented out, e.g. *esnoBAR --> <stress> esnobar*

4. Introducing "dishesion marker candidates" (eee)

* Due to a complete lack of full stops, colons and commas (only question marks [?] and turn taking [¶] are used), other means of marking syntactic windows become necessary, and strings like ..., eh, éh, () are marked as "dishesion elements", as well as quotes if they enclose more than 1 word.

5. Mapping "dishesion markers" as

a) <break> (major syntactic break, clause or sentence boundary)

* <break> markers can be used by the CG rules to establish maximal group size or valency scope; e.g., <break> should not occur between a premodifier and its head, or between main verb and direct object.

b) <pause> (non-word hesitation/pause marker)

* <pause> markers are not allowed to break up group og clause continuity.

6. Disambiguating "dishesion markers" (eee)

a) "xxx" --> eee xxx eee --> <pause> xxx <pause>
If a single word is surrounded by dishesion markers, these are treated as <pause>

b) eee (e) que/quando/embora ... --> <pause>
If a dishesion marker is followed by a conjunction or relative, possibly with an interfering coordinator, it is treated as <pause>.

c) que/quando/embora ... eee --> <pause>
If a dishesion marker is preceded by a conjunction or relative, it is treated as <pause>

d) eee + PRP --> <pause>
If a dishesion marker is followed by certain prepositions (de, em, com, sem, por), it is to be treated as <pause>

e) PRP/det + eee + NON-art/dem --> <pause>
If a dishesion marker is preceded by a preposition or a determiner (or a fused presposition+determiner), it is to be treated as <pause>, unless it is directly followed by an article or demonstrative (in which case the <pause>/<break> ambiguity is retained)

4. Grammar tasks

The next effort concerned the CG rule grammar as such, where dishesion marker candidates had to be integrated in those tag sets that denote possible syntactic breaking points. The PAUSE set, for example, includes not only the dishesion marker, but only certain interjections:

LIST PAUSE = "uhn" "ahn" "eh" "eee" <pause> <break> IN ;

The <break> tag is useful in NON-sets since these are often used in BARRIER conditions in CG-rules, baring group attachment, for instance:

LIST NON-NP = PERS SPEC ADV VFIN INF PRP KS KC <rel> <interr> "<$\,>" <break> >>> <<< ;

On the sentence level, <break> is a potential clause boundary marker, the same way certain complementizers, comma and hyphen are:

LIST CLB = KS <interr> <rel> "<$\,>" "<$->" KOMMA <break> ;

Next, rules have to be crafted for the disambiguation of cohesion markers - are they breaks denoting "sentence" window borders, or just pauses embedded in the syntactic flow of speech?

For instance, dishesion markers are not <break> (but <pause>) if they intervene:

(a) between a "name bearer" and its name: *o rei $$ Alfonso*
(b) between a noun and the preposition 'de': *pai $$ de muitos filhos*
(c) between an intensifier and an attribute: *uma maneira um pouco $$ calcada*
(d) between a noun and a potential postmodifier or object complement of the same gender and number: *estou vendo **a TV** evidentemente $$ muito **presa** a ...*
(e) between a transitive main verb and its direct object.

(a) REMOVE (<break>) (-1 (<+n>)) (1 <*>)
(b) REMOVE (<break>) (-1 N) (1 PRP-DE)
(c) REMOVE (<break>) (-1 <quant>) (1 ATTR/<attr>)
(d) REMOVE (<break>) (*-1 NFP BARRIER ALLuPAUSE/ADV) (*1 ATTR-FP BARRIER ALLuPAUSE/ADV)
(e) REMOVE (<break>) (-1C @MV LINK 0 <vt>) (*1C @<ACC BARRIER @NON->N)

Of course, the use of dishesion markers and their introduction in NON-sets and CLB-sets, has to be balanced between the advantages of providing better defined windows of analysis, and the draw-backs of disallowing many long range rule contexts that are conditioned by CLB-barriers and the like.

5. Special problems

a) premodifier clashes (da dos)

In a simple correctional article clash ('comeu a o bolo') both articles will receive the @>N (premodifier) tag, but in more complex cases there may be problems, for instance, where a

preposition is repeated as well. Here, the first determiner will be analysed as @P< (argument of preposition).

eu	não	estou	agora	por	dentro	**de**	**a**	**de**	**os**	nomes	sabe	?
SUBJ>	ADVL>	FMV	ADVL>	<SC	P<	A<	**P<**	N<	>N	P<	FMV	

b) "faulty" noun phrases: stranded premodifiers in incomplete np's (um, uma) **and agreement errors** (codificação nada normativo)

* In speech more than in text, a long distance between head and modifier may result in agreement lapses (here, M - F).

* stranded premodifiers tend to assume np-head function in a syntactic parse, which may seem odd, but is hard to avoid, and may well be the logical solution - after all, in a word-based tagger/parser there are no zero constituents, and every function has to be attached *somewhere.*

e	$e	não	havendo	uma	codificação	não	$pause
CO		ADVL>	IMV ICL-ADVL>	>N	<ACC	ADVL>	

$break	$eee	**um**	**uma**	$pause	nada	**normativo**
		<ACC	<ACC		>A	N<

Agreement failure (here SG - PL) does occur in adjacent position, too. The examples are taken from a transscription where the speaker (a lecturere) admitted to being nervous on being taped.

a	demanda	de	moeda	por	transação	$pause	é	$pause	
>N	SUBJ>	N<	P<	N<	P<		FMV		

principal	**motivo**	por	**os=quais**	as	pessoas	$pause	retêm	moeda
>N	<SC	ADVL>	P< FS-N<	>N	SUBJ>		FMV	<ACC

nós	podemos	resumir	isso	**em**	**um**	**exemplinhos**	numérico
				<ADVL	>N	P<	N<

c) true ambiguity with regard to primary sentence constituents

estão	gravando	agora	este	$pause	está	passando	$está passando
FAUX	IMV ICL-AUX<	<ADVL	<ACC		FAUX	IMV ICL<AUX	

agora	em	São=Paulo	O	Grito	não	é	?

<ADVL	<ADVL	P<	>N	<SUBJ <ACC <SC	ADVL>	FMV

Note that the stranded premodifier ‘este’ recieves the funtion tag of its presumed np-head.

d) difficulties in identifying subjects:

Consider the following example, where three subject tags have to be found and tolerated in the same speech chunk without clear clause boundaries:*televisão, ela, telespectador:*

porque	a	**televisão**	sendo	estatal	**ela**	é	muito	$stress
SUB FS-<ADVL	>N	SUBJ>	IMV ICL-<ADVL	<SC	SUBJ>	FMV	>A	

uniformizada	$pause	$break	não	há	espectáculos	diversificados	o
<SC			ADVL>	FMV	<ACC	N<	>N

telespectador	$pause	$break	**o**	fica	sempre	$pause	preso
SUBJ> <ACC			ACC>	FMV	<ADVL		<SC

a	filmes	ou	a	$a	conferências		
A<PIV	P<	CO	<PIV		P<		

Here, ‘**ela’** is semantically anaphoric to ‘televisão’, which syntactically belongs to its own non-finite subclause. ‘**telespectador’** lacks a sentence/analysis window marker (before its article), which is why function has not been fully disambiguated in this case.‘**o**’ before the main verb ‘fica’ might be part of yet another subject candidate with only its article left, but since the grammar strongly disallows adjacency of articles and finite verbs, ‘o’ is treated as a personal pronoun in the accusative. ‘o’ does not bear any meaning in this sentence, and would be ignored by a human listener, but once uttered and transscribed, the word has to be handled in the grammar one way or another.

e) notationally caused and interaction based errors in multi-speaker data

In a notation that uses only one time line, utterances of speaker S2 may syntactically “cut” an utterance of speaker S1. Also, speakers S1 and S2 may interact syntactically, finishing each others groups or clauses. In the example, ‘adequado’ (S1) is subject complement (SC) for ‘está’ (S2), ‘perfeitamente’ (S2) is premodifier (>N) for ‘adequado’ (S1):

*L2 para aquele ... está **perfeitamente** ...*
*L1 **adequado***
*L2 **adeQUAdo:: do** ... é muito mais interessante ... é uma*
[
L1()
L2 grande oportunidade para os nossos artistas nåo é ?
L1 isso é muito bom:: eh:: e ain/ e:: e a novela puxa o disco porque parece que na vendagem dos discos eles såo muito ... requisitados esses discos de novelas né ?
L2 H. você escreveu qualquer coisa muito interessante

6. Positive side effects: robustness

CG's flat analysis is very robust, and especially advantageous with unclear sentence boundaries or nested sentences, both of which are frequent in speech data. Consider the 5 main verbs in the following comma- and coordinator-free sentence:

e	**é**	uma	grande	atriz	$break	então	**choca**	demais	$pause	$break
CO	**FMV**	>N	>N	<SC		ADVL>	**FMV**	<ACC		

aquela	paulista	$stress	quatrocentona	**que**	ele	**faz**	bem	$stress
>N	<SUBJ		N<	ACC> FS-N<	SUBJ>	**FMV**	>A	

grifado	$break	aliás	de	uma	maneira	um=pouco	$pause	calcada
<OC		ADVL>	ADVL>	>N	P<	>A		N<

demais	**porque**	esse	tipo	**acho**	**que**	já	se	**diluiu**
A<	SUB	>N	SUBJ>	**FMV**	SUB FS-<ACC	ADVL>	ACC>	**FMV**

nem	**existe**	mais	$pause	mas	...
<ADVL	**FMV**	<ADVL		CO	

Even double main verbs without <u>any</u> sensible syntactic analysis, and breeches of the uniqueness principle are tolerated fairly well by the CG-grammar:

$break	isto	**é**	**levava**	a	um	tipo	de	vida	nômade
	SUBJ>	**FMV**	**FMV**	<PIV	>N	P<	N<	P<	N<

When all goes well, the system tolerates overlapping clauses with double uncoordinated subjects and a shared direct object, as well as - to a certain degree - complex and interrupted np's and np-modifiers (boxes).

		problems:
papai	N M S @SUBJ>	
mesmo	DET M S @N<	
tem	V PR 3S IND @FMV	obligatorily transitive verb *without* direct object
em	PRP @<ADVL	
os	DET M P @>N	
<$nos>		
livros	N M P @P<	

de	PRP @N<	
ele	PERS M 3S NOM/PIV @P<	
ele	PERS M 3S NOM/PIV @SUBJ>	2 subjects without co- or subordination
tem	V PR 3S IND @FMV	2 main verbs without co- or subordination
muitas	DET F P @>N	
expressões	N F P @<ACC	direct object serving verbs in 2 clauses
$pause		
completamente	ADV @>A	
caídas	V PCP F P @N<	heavy postnominal with adjunct and argument
em=desuso	VPP @A<PIV	
e	KC @CO	
portuguesas	**N** F P @<ACC**??**	less heavy postnominal *after* heavy postnominal
e	KC @CO	
<$por/>		
e	KC @CO	
$pause		
de	PRP @SC> @N<	very distant pp-postnominal with false start
português	N M S @P<	
clássico	ADJ M S @N<	
não	ADV @ADVL>	
é	V PR 3S IND @FMV	finite clause without punctuation or header
$?		

7. Performance

A quantitative comparison of the two versions of the parser (written language vs. speech data) yielded the following results, with correctness defined as *recall at near 100% disambiguation*, counting both false tags, missing tags and false ambiguity as errors.

Parser performance on running text (VEJA news magazine and fiction)

Text:	*O tesouro* ca. 2500 words		*VEJA 1* ca. 4800 words		*VEJA 2* ca. 3140 words	
Error types:	errors	correct	errors	correct	errors	correct
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	
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 %**

Parser performance on speech data (before/after grammar adaptation)
(NURC *[norma lingüistica urbana culta]*, São Paulo)

speech sample	sample size	morphological correctness	syntactic correctness
2 speaker dialogue (topic: cinema, television, actors) females, 60 yrs (journalist and writer)	2810 words	99.2 %	95.7 %
secondary school teaching monologue (history), female 36 yrs	2080 words	99.5 %	96.3 %
university teaching monologue (economics), male 31 yrs	1600 words	99.0 %	95.4 %
base line: 2 speaker dialogue (same as above) analysed with unmodified grammar	1100 words	98.9 %	92.6%

8. Conclusion

While an automatic parser originally designed for written Portuguese was able to more or less maintain its performance on speech data *morphology* (word class etc.), error rates tripled for speech data *syntax*. Judging from the effectiveness of according rule changes and preprocessing, one can conclude that at least one of the reasons for this striking difference resides in the fact that the disambiguation of morphological ambiguity involves mostly short range group context that is left intact even in the grammatically often incomplete utterances of spoken language, while rule based syntactic analysis depends on long range context patterns, working less than perfect without a clear sentence window, without full complementation of obligatory valency, and with breaches of the uniqueness principle. The hypothesis was tested by tagging - through a preprocessor module - what I call *dishesion markers* ("...", "eh" etc.) in the corpus as both <pause> and <break> for later disambiguation, thus introducing "sentence boundary" candidates, which may be disambiguated by either crude word form context or elaborate long range CG rules. Once disambiguated, the <break> markers provide more "traditional" syntactic window delimiters for the system's Constraint Grammar, considerably improving syntactic tag recall. Examples where modification of the syntactic rules as such proved necessary are violations of the uniqueness principle due to iterations or modified ("corrected") iterations, or cases, where one speaker complements the valency pattern of a syntactic unit uttered by

another speaker. Especially problematic are clashes, where a speaker strands dependents without their heads (for instance, subjects without a verb, or a premodifier without its nominal head) and departs on a new syntactic path.

Preliminary quantitative results suggest that break markers and rule modifications can narrow the gap between the parser ‘s performance on written and spoken Portuguese, respectively, to a few percentage points (i.e. 95-96% correctness) for syntax and nearly eliminate it for part of speech tagging.

References:

Bick, Eckhard, 1996. “Automatic Parsing of Portuguese", in *Proceedings of the Second Workshop on Computational Processing of Written Portuguese,* Curitiba:CEFET-PR.

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

Bick, Eckhard, 1997. "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:Institut for erhvervssprog og sporglig informatik, Handelshøjskole Syd.

Birn, Juhani, 1998. “Swedish Constraint Grammar”, in *Proceedings of the 17th Scandinavian Conference on Linguistics*, this volume. Odense.

Castilho, Ataliba Teixeira de (ed.), 1989. *Português culto falado no Brasil*. Campinas: Editora da Unicamp.

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

Hagen, Kristin and Janne Bondi and Anders Nøklestad, 1998. “A Constraint-Based Tagger for Norwegian”, in *Proceedings of the 17th Scandinavian Conference on Linguistics*, this volume. Odense.

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