Automatic Grammatical Annotation of Historical Brazilian Portuguese

Eckhard Bick

University of Southern Denmark

eckhard.bick@mail.dk

Outline

- Introduction: Annotation of historical Portuguese
- Corpora: Reference corpora, Letters & Editorials, Colonia corpus
- Automatic grammatical annotation
  - PALAVRAS parser setup
  - adaptations for historical data
- Evaluation
- Creating a diachronic dictionary
  - dictionary formats
  - results
- Conclusions & Outlook

Part 1:
Annotation of historical (Port.) corpora

- Historical texts are difficult to handle with language technology
  - material: hand-written, OCR
  - bibliographical meta data and comments may be in-text
  - language: non-standard orthography, lack of standardization, archaic lexicon and grammar
- Research objective: Under these conditions, can an existing NLP system be modified for historical data? How?
- Resource objectives:
  - (a) Linguistically annotate raw historical corpora, enrich existing annotation (syntax, semantics)
  - (b) Generate an on-the-fly dictionary of diachronic variation in Portuguese for a specific (sub)corpus, mapping spelling variation in a particular period, author or text collection
- Perspective: method resuable for other non-standard spelling: e.g.
  - speech transcriptions with phonetic modifications
  - social media jargon
  - learner language ...

Reference corpora of Historical Portuguese

- Tycho Brahe Corpus (Paixão de Sousa and Trippel, 2006; Galves and Faria, 2010)
  - syntactic focus
- HDBP project (Candido and Aluísio, 2009)
  - lexicographical focus, Brazilian
- Corpus do Português (Davies, 2006, 2014)
  - 45 M words, European and Brazilian genre & historical
- GMHP (Universidade de São Paulo)
  - focus on morphology
- Colonia (Zampieri & Becker, 2013)
  - 5 M, mixed period & mixed variety

São Paulo Letters & Editorials Corpus (early Brazilian Portuguese)

- original paper documents, ca. 121.000 tokens (Barbosa & Lopes 2002)
- 19th century, regionally homogeneous
- philological sources converted into a text corpus
  - meta text, footnotes etc. kept separate from the corpus proper
  - reconstitution of "broken" words
    - line break hyphenation
    - manually marked breaks '|'
- Syntactically motivated tokenization
  - splitting of historically fused expressions (12% increase in token count, 3rd colum), semiautomatical
    - e.g. coordinator/preposition + noun, clitic + verb
  - fusion of names into MWE tokens ("functional words"), automatical

Corpus size and word/token distribution

	Period	Size (annotated words)	**Size (functional words)**	Size (token words)
aldeamentos de índios	1722-1809	12.951	11.853	11.215 (- 13%)
cartas paulistas, biblioteca RJ	1801-1822	16.513	14.935	14.433 (- 13%)
Cartas, cap.6 (Barbosa & Lopes)	1827-1900	36.755	33.774	33.457 (- 9%)
Anúncios (Guedes & Berlinck)	1829-1899	64.477	55.787	57.910 (- 13%)
correspondência Washington Luiz	1897-1900	4.387	4.040	4.076 (- 7%)
All	1722-1900	135.083	120.389	121.091 (- 10.4%)

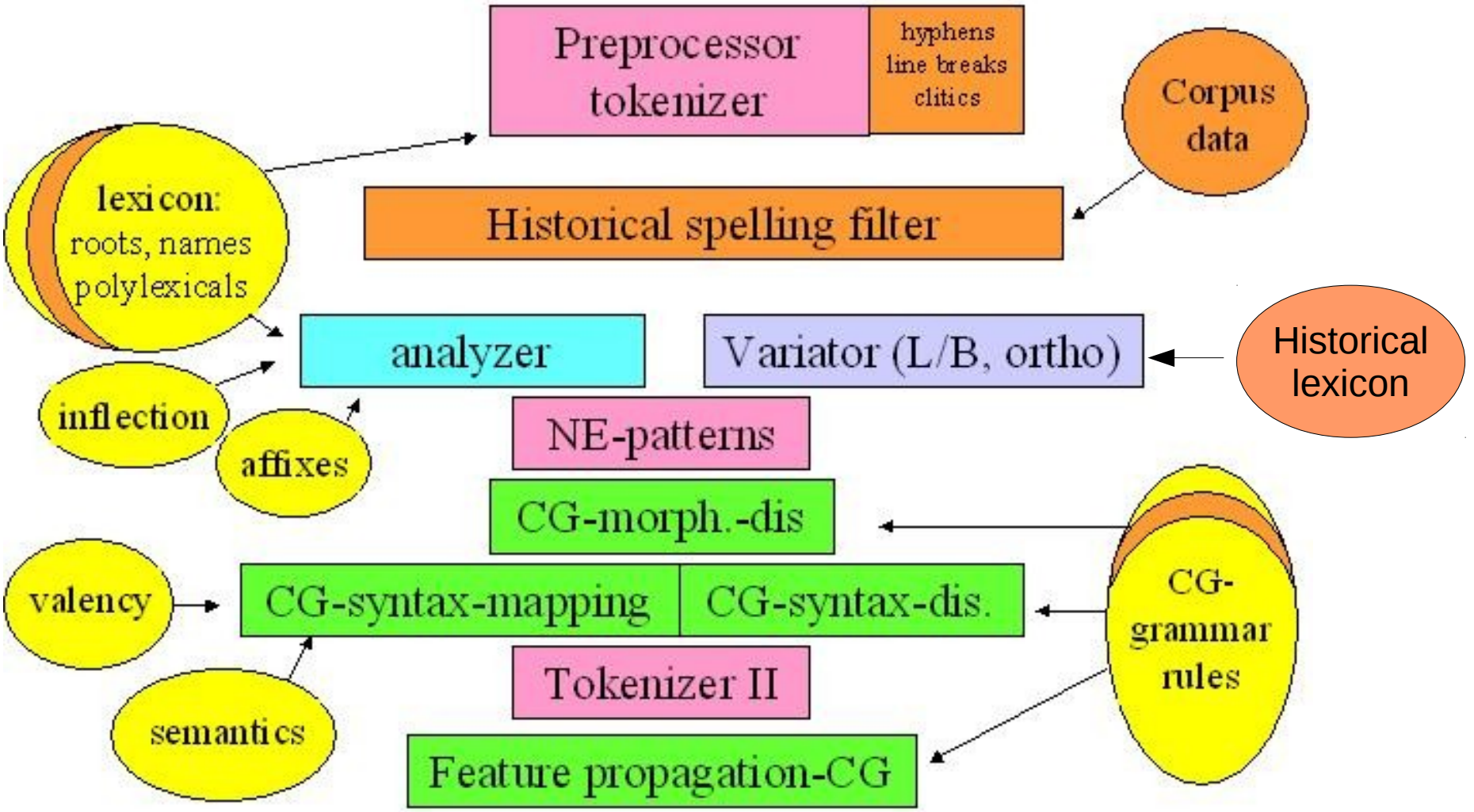
Automatic grammatical annotation

- add linguistic value on top of philological mark-up, e.g.
  - lemmatization --> lexicography
  - syntax --> diachronic changes in constituent order, valency patterns
- hand-annotation is very time consuming
- training a parser is difficult
  - lack of training data
  - problematic tokenization: "prosodic" word fusion
- proposed solution: adapt a rule-based parser
  - rule-based parsers don't need training data and are therefore less corpus/domain-specific and less sensitive to language variation, including historical data
  - rule-based parsers allow transparent and specific interference by a linguist
  - but the parser will need either
    - a historical lexicon or
    - orthographical "translation", or both

The PALAVRAS parser

- robust, rule-based system handling both European and Brazilian Portuguese
- earlier experiments with non-standard data
  - dialectal: Cordial-Syn
  - speech: NURC, C-ORAL Brasil
- earlier experiments with historical data: Tycho Brahe
- PALAVRAS has a suite of postprocessing tools
  - constituent tree format, MALT xml, TIGER xml, CoNLLformat, UD treebank format, ...

System architecture



PALAVRAS' annotation scheme

- standard fields: (1) Word - (2) [lemma] -
  (3) <secondary tags> - (4) POS - (5) inflexion -
  (6) @syntax/function - (7) dependency relations
- added fields <OALT:...> for normalised word form

Esta 'this'	[este] <dem>	DET F S	@>N #1->2
povoaçam 'settlement'	[povoação] <OALT:povoação> <Lciv>	N F S	@SUBJ> #2->3
he 'is'	[ser] <OALT:é>	V PR 3S IND	@FS-STA #3->0
uma 'a'	[um] <arti>	DET F S	@>N #4->5
Villa 'town'	[vila] <OALT:Vila> <Lciv>	N F S	@<SC #5->3
mui 'very'	[muito] <OALT:muito> <quant>	ADV	@>A #6->7
fermosa 'famous'	[fermoso] <ORTO:formosa>	ADJ F S	@N< #7->5

Dependency trees

A	[o] <artd> DET F S	@>N	#1->2
expedição	[expedição] N F S	**@SUBJ>**	#2->6
contra	[contra] PRP	@N<	#3->2
o	[o] <artd> DET M S	@>N	#4->5
Mexico	[México] PROP M S	@P<	#5->3
sahio ALT saiu	[sair] <fmc> V PS 3S IND	**@FMV**	#6->9
a	[a] PRP	**@<ADVL**	#7->6
5	[5] <card> NUM M/F P	@P<	#8->7
de	[de] PRP	@N<	#9->8
Julho	[julho] N M S	@P<	#10->9

Tree structures

STA:fcl

SUBJ:np

=>N:art('o' <artd> M P) os

=H:n('soldado' M P) soldados

=N<:fcl

==SUBJ:pron-indp('que' <rel> M S) que

==ACC:pron-pers('ela' F 3S ACC) a

==P:v-fin('compor' PR 3S IND) compõe

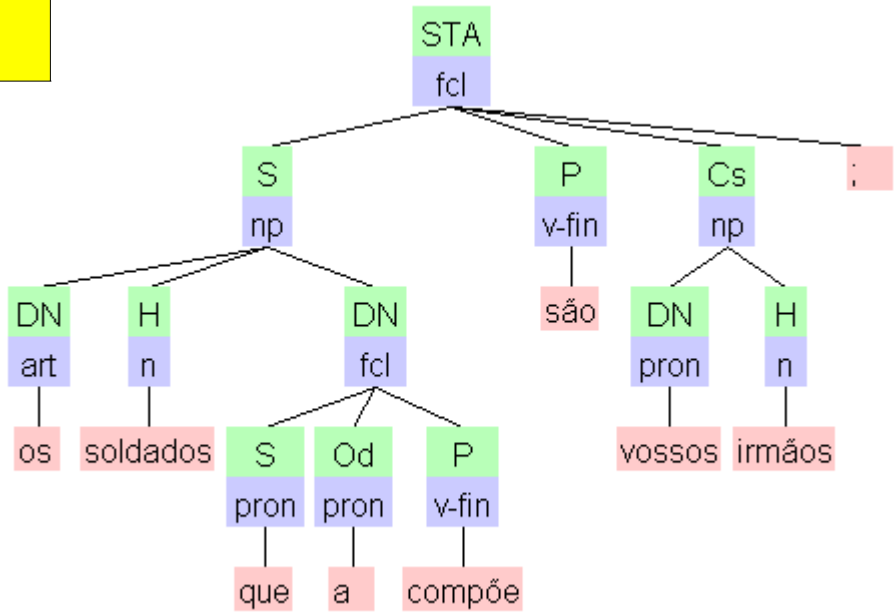
P:v-fin('ser' PR 3P IND) são

SC:np

=>N:pron-det('vosso' <poss> M P) vossos

=H:n('irmão' M P) irmãos

CG constituent format



VISL treebank format

A problem for the parser: Examples of historical variation

- geminated and triple consonants
  - dd->d, ff->f, cc[aou]->c, sss->ss etc. *(attenção, accumula, soffra, affligir)*
- word fusion
  - *heide, hade -> hei de, há de*
- "Greek"/classical spelling
  - ph->f, th->t, y -> i *(mathematica, authores, systema)*
- nasals
  - em[dt]->en *(bemdito),* om[df]->on *(comforme),* aon->ão *(christaons)*
  - chaotic -ão/am and -ões: *áo, ào, âo, aõ, aò, àm, ao*, *ôes, óes, oens,* ans -> ãs *(irmans)*
- others:
  - extra hiatus-h: *sahiu, incoherente, comprehender*
  - z/s-dimorphism: isa -> iza, [aeu]z -> s, [óú]s$ -> z *(civilisadas, acuza, uzo, brazileiros, cazo),* cave: *crus (->z), gosa (->z), cafuzo, avós, após*
  - s/c-ellision: sci -> ci, cqu -> qu: descifrada*, sciência*
  - lack of tonic accents: *anivers**a**rio, mal**i**cia, razo**a**vel, provid**e**ncia, far**i**amos*
  - "superfluous accents": *d**ó**e, pess**ô**a*
  - fluctuating accents: *n**ò**s, ser**à**, judа**ï**smo*

Orthographical standardization

- rule-based with regular expressions (Hirohashi 2005)
  - Tycho Brahe: experiments with tagger lexicon extensions first
  - HDBP (Historical Dictionary of Brazilian Portuguese): lumps variants around a common "base form", but not necessarily the modern one (Candido & Aluísio 2009)
- statistical spelling normalization: VARD2 for Portuguese (Hendricks & Marquilhas 2011)
  - recovers 61% of variations, 97% of which point to the correct standard form
  - normalisation improved subsequent POS tagging
- neural-network-learned, post-edited POS tags *before* morphological analysis (Rocio et al. 2003)
  - hand annotation of 10,000 words per text *without* normalization
  - followed by 250 DCG rules for partial parsing of modern Portuguese
- Our approach: rule-based normalization, as modern as possible
  - why? (1) no need for hand-annotating training data
  - Why? (2) normalisation into modern forms allows the use of standard parsing tools

Parser adaptations 1

- Preprocessor: recognition of fused word parts, function word heuristics for unknown forms
  - Spanish-style clitics
  - fused prepositions: *dasua, daSua*
  - articles and conjunctions: *eogrande*
  - apostrophed vowel ellision: *ess'outra*
- Language filter recognizing Latin, Spanish, French, Italian and English segments and blocking them from Portuguese analysis
  - word-based voting system, Portuguese-biased thresholds
  - necessary, because orthographical "relaxation" for individual words would lead to many false positives in e.g. a Portuguese lemma inventory
- Historical spelling filter
  - standardizing historical letter combinations and inflection paradigms
  - 2-level annotation, where the original form is stored, while the standardised form is used by the parser

Parser adaptations 2

- Fullform lexicon of modern word forms
  - generated from the parser lexicon by applying inflection paradigms (ca. 500,000 entries)
  - used to validate/constrain standardisation candidates (avoid false positives) and for restoring/changing accents
    - closed/open vowel marking: *recebêram, levára, chóra*
    - orthographically expressed phonetic variation, e.g. 3.ps. PS: *consentio, envadio, attrahio, commetteo, encheo*
    - plural variants: *officiaes, quaes*
- External dictionary and morphological analyzer, supplementing the parser's own morphological module
  - adds (historical) readings to the (heuristic) ones used by the parser for unknown words
  - letting contextual Constraint Grammar rules decide in case of POS ambiguity
  - also used for the numerous Tupi and other regional words in the corpus, to prevent them from getting "standardised" into similar, but wrong modern words
- Remaining problems
  - False negatives, where a word form is thought to be modern, but in fact should have been changed
    - *vera* ADJ? - *verá* V FUT
  - Ambiguity: *obrigaÇam* --> *obrigassem* instead of *obrigação*

Evaluation of orthographical filtering on the Letters & Editorials annotation

	% correct PoS & morphology	% correct syntactic function	tokenization errors	"sentence" length (words)
(1) aldeamentos de índios	95.4 %	91.5 %	1.7 %	~ 60
(2) cartas paulistas, biblioteca RJ	95.5 %	90.7 %	1.1 %	~ 26
(3) Cartas, cap.6 (Barbosa & Lopes)	98.2 %	94.3 %	0.0 %	~ 22
(4) Anúncios (Guedes & Berlinck)	97.0 %	92.0 %	0.5 %	~ 14
(5) correspondência Washington Luiz	97.2 %	92.6 %	0.0 %	~ 21
Average	**96.7 %**	**92.2 %**	**0.7 %**	**28.6**
Modern Portuguese (mixed genre)	> 99 %	> 96 %	-	-

- tokenization problems and sentence length correlate inversely with tagging accuracy

E porque ao serviço deDeos, e de Sua Magestade e boa admninistraçaõ dos mesmos indios he conveniente dar a Vossa Excelencia plena informaçaõ destas aldeas sou obrigado a manifestar que das aldeas que actualmente administramos nenhuâ he das que se chamaõ nesta terra aldeas de Sua Magestade por que estas sendo antigamente de gente innumeravel fundadas pellos Religiozos da Companhia fomos obrigados a dimiti-las de noSso governo cançados de as naõ podermos defender dos injustos cativeiros de homês poderozos; (80 words)

Orthographical variation quantified across the time axis

	e-fusion	all non-clitic fusion	orthographical heuristics/lexicon	old h-words (not initial, nh, lh, ch)
aldeamentos 1722-1809	102/454 22.5 %	232/4215 5.5 %	2.908 24.5 %	41 0.3 %
cartas paulistas, bibl. RJ, 1901-1822	274/571 48.0 %	488/5610 8.7 %	2.840 19.0 %	61 0.4 %
Cartas de leitores e redatores, 1827-1900	2/936 0.2 %	34/13509 0.3 %	4.228 12.5 %	153 0.5 %
Anúncios 1829-1899	4/1623 0.2 %	56/25296 0.2 %	6.694 12.0 %	387 0.7 %
correspondência W.Luiz, 1897-1900	0/129 0.0 %	5/1610 0.3 %	484 12.0 %	71 0.2 %
Average	0.3 %	1.1 %	16.0 %	0.4 %

subject/object percentages in the weighted revised corpus

	of all @		N/PROP	PERS
@SUBJ>	4.6 (5.6) [PE 5.5 - PB 6.8]	of these:	46.1 (52.0) [PE 69.0 - PB 74.1]	**16.7** (14.2) [PE 6.0 - PB 8.3]
@<SUBJ	**1.3** (1.0) [PE 0.8 - PB 0.7]	of these:	66.7 (73.9) [PE 80.2 - PB 86.2]	**13.3** (17.4) [PE 12.9 - PB 7.2]
@ACC>	**1.8** (2.1) [PE 0.9 - PB 0.8]	of these:	0.1 (0.0) [PE 0.3 - PB 0.6]	55.0 (70.0) [PE **50.5** - PB 57.3]
@<ACC	4.6 (4.4) [PE 4.5 - PB 4.9]	of these:	68.9 (69.7) [PE 83.9 - PB 90.8]	13.6 (13.6) [PE 9.7 - PB **2.7**]

PE = modern European Portuguese
PB = modern Brazilian Portuguese
(...) = category frequency in the unrevised annotation

* OV and VS word order is rare/marked also in historical Portuguese, but less so (f = x 2)
* post-positioned object clitics: Historical PB similar to modern PE

Part 2:
Creating a historical dictionary for Portuguese

- philological considerations (source, period, author, typeset, ...
- manual vs. automatic compilation
- dictionary-based or corpus-based
- historical root dictionary (Silvestre & Villalva 2014)
  - lexical analysis, etymology, other dictionaries as source
- corpus-based philological dictionary (HDBP, Murakawa 2014)
  - definitions and quotations for historical usage
  - a) published version lumping variants under 10,500 modern-spelled entries
  - b) automatically extracted glossary of 76,000 variants of 31,000 "common" forms
  - c) manually compiled dictionary of 20,800 token fusions ("junctions)
- So why yet another resource?

Dictionary format, comparison

HDBP automatic glossary	Colonia corpus dictionary
Brazilian only	cross-variant, potentially broader focus
does not resolve POS, no inflexional analysis --> difficult to extract category-based patterns	parsing-based, morphological analysis and contextual disambiguation
no period/author differentiation	possibilities for on-the-fly subdictionaries for periods/authors
modern and historical entries are mixed (villa -> vila, but tão -> tam, chamam -> xamam, também -> tambem, tãobem separate)	systematical use of modern forms as standard, *even with automatic extraction*
strips acute and circumflex accents, creating ambiguity even in modern forms (contínua ADJ vs. continúa/continûa/continúa V) '-ão' is not disambiguated when meaning '-am' (matarão - mataram)	all forms are shown as is, and linked to a modern standard form
the glossary contains unmarked fusions (foime), and doesn't seem to use the separate junction lexicon	fusions are automatically split

The Colonia Corpus

- complete Portuguese manuscripts published 1500-1936
- 5 subcorpora per century, variety-balanced with 48 pt - 52 br

Century	Texts	Tokens
16th	13	399,245
17th	18	709,646
18th	14	425,624
19th	38	2,490,771
20th	17	1,132,696
Total	100	5,157,982

- first version distributed with treetagger POS annotation
- 5,1 M tokens compiled from different repositories
  - Tycho Brahe (Galves & Faria 2010)
  - GMHP corpus
  - Domínio público database
- used for various NLP tasks and linguistic studies, e.g.
  - temporal text classification (Niculae et al. 2014)
  - style variation and stylometrics (Ŝtajner and Zampieri 2013)
  - diachronic morphology (Tang & Nevins, 2013))
  - lexical semantics (Santos and Mota 2015)

Evaluation of the effect of orthographical filtering

Century	Words	Treetag. unknown (- PROP)[5]	PAL, heuristic lemma (- PROP)	Treetag. accuracy (POS)	Accuracy modified PAL. (POS)	Accuracy mod. PAL. (syntactic function)
16th	473	15.2	0.4	80.1	96.6	91.1
17th	432	0.7	0.0	96.5	98.8	94.4
18th	477	21.8	0.6	81.1	97.7	91.6
19th	372	1.3	0.8	95.2	98.1	93.3
20th	446	0.2	0.0	97.3	99.6	96.0

- The modified PALAVRAS outperforms the original Colonia tagging (Treetagger)
- performance decrease for older texts is buffered by orthographical filtering
- correlation between lexical coverage and tagging accuracy
- exeption of age-accuracy correlation: 17th century (newspaper sources with a likely high level of standardization/proof-reading)
- moderate decrease in syntactic performance (only indirectly affected by orthographical filtering, no specific rules for e.g. VS, VOS, OVS word order)

Generating a diachronic dictionary

- Extraction of all normalized forms and split parts of fusions
- Exclusion of foreign language forms (chunk size > 3)

Century	Words	Orthographically non-standard	Fused	Fused (relative)
16th	528K	4.11 %	0.93 %	22.63 %
17th	577K	2.09 %	0.25 %	12.31 %
18th	456K	2.88 %	0.25 %	8.68 %
19th	2,459K	0.30 %	0.08 %	28.23 %
20th	857K	0.17 %	0.04 %	23.52 %

Century	All Foreign	Latin	Spanish	Italian	French
16th	0.78 %	0.49 %	0.26 %	0.01 %	-
17th	0.78 %	0.51 %	0.24 %	-	0.01 %
18th	0.23 %	0.17 %	0.05 %	-	-
19th	0.03 %	0.01 %	-	-	0.01 %
20th	0.03 %	0.01 %	0.01 %	-	0.01 %

Dicionary size and layout

- 10,400 wordform types, representing 52,000 corpus tokens, with distribution across century periods
  - capitaes <OALT:capitães> (1; 16th:1 - - - -)
  - capitaina <ORTO:capitânia (5; 16th:5 - - - -)
  - capitam <OALT:capitão> (5; 16th:4 - 18th:1 - -)
  - capitan <OALT:capitão> (1; 16th:1 - - - -)
  - capitanîas <OALT:capitanias> (1; - 17th:1 - - -)
  - capitaõ <OALT:capitão> (3; - 17th:1 18th:2 - -)
  - capitaens <OALT:capitães> (14; - 17th:10 18th:4 - -)
- 862 non-standard word fusion types, representing 5,000 tokens
  - *ess'outra ("this other")*
  - *fui-lh'eu ("I was [for] him")*
  - *estabeleceremse ("to establish themselves 3.pl.")*

Conclusion and outlook

- results
  - given orthographical standardization, a standard rule-based parser can achieve reasonable performance across a wide range of historical texts
  - standardization was most important for the 16th-18th century
  - our method can produce tailor-made historical wordform dictionaries
  - our method provides a solution to the out-of-vocabulary problem encountered by statistical taggers when used on historical text
- problems
  - false negatives, where a word matches an existing modern form, but still should have been changed (eg. *noticia V?* vs *notícia* N)
  - ambiguous substitutions (*estillo* -> *estilho?* [Spanish ll->lh] *vs.* -> *estilo* [gemination variant]
- future work
  - focus on syntactically motivated grammar modifications on top of orthography

eckhard.bick@mail.dk

Parser:
http://visl.sdu.dk

Corpora:

http://corp.hum.sdu.dk/cqp.pt.html
http://corporavm.uni-koeln.de/colonia/interface.html

References

Bick, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. In: Working with Portuguese Corpora. London/New York:Bloomsburry Academic. (2014) 279-302

Bick, E., Módolo, M. Letters and Editorials: A grammatically annotated corpus of 19th century Brazilian Portuguese. In: Romance Corpus Linguistics II: Corpora and Historical Linguistics (Proceedings of the 2nd Freiburg Workshop on Romance Corpus Linguistics, Sept. 2003). Tübingen: Gunther Narr Verlag. (2005) 271-280

Bick, Eckhard & Marcos Zampieri. Grammatical Annotation of Historical Portuguese: Creating a Corpus-Based Diachronic Dictionary. In: Sojka, P. & A. Horák & I. Kopeček & K. Pala (eds.), Text, Speech and Dialogue - 19th International Conference, TSD 2016 (Brno, 12-16 Sept 2016). LNAI Series, Vol. 9924. Heidelberg: Springer (2016). pp. 3-11

Candido, A., Aluísio, S. M.. Building a Corpus-Based Historical Portuguese Dictionary: Challenges and Opportunities. In: TAL 50(2). (2009) 73-102

Davies, M. Creating and Using the Corpus do Português and the Frequency Dictionary of Portuguese. In: Working with Portuguese Corpora. London/New York:Bloomsberry Academic. (2014) 89-110

Galves, C., Faria, C. Tycho Brahe Parsed Corpus of Historical Portuguese. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html. (2010)

Hendrickx, I, Marquilhas, R. From old texts to modern spellings: an experiment in automatic normalisation. In: Journal for Language Technology and Computational Linguistics, 26(2). (2011) 65-76

Hirohashi, A. S. Aprendizado de regras de substituição para normalização de textos históricos. Master Thesis – Institute of Mathematics and Statistics, USP, São Paulo, Brazil. (2004)

Murakawa, C.A.A. A construção de um dicionário histórico: o caso do Dicionário Histórico do Português do Brasil — séculos XVI, XVII e XVIII. In: Estudos de lingüstica galega 6. (2014) 199-216

Niculae, V., Zampieri, M., Dinu, L. P., Ciobanu, A. M. Temporal text ranking and automatic dating of texts. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL). (2014) 17-21

Rocio, V., Alves, M.A., Lopes, J.G., Xavier, M.F., Vicente, G. Automated Creation of a Partial Portuguese Treebank: Building and Using Parsed Corpora. In: Abeillé, A. (ed) Treebanks. Vol. 20 of Text, Speech and Language Technology. Dordrecht: Springer. (2003) 211-227

Santos, D., Mota, C. A admiração à luz dos corpos. In: Oslo Studies in Language 7.1. (2015) 57-77

Silvestre, J.P., Villalva, A. A Morphological Historical Root Dictionary for Portuguese. In: Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen. Bolzano/Bozen: EURAC research. (2014) 967-978

Štajner, S., Zampieri, M. Stylistic Changes for Temporal Text Classification. Proceedings of Text, Speech and Dialogue (TSD), Lecture Notes in Artificial Intelligence - LNAI 8082, Springer. (2013) 519-526.

Paixão de Sousa, M.C., T. Trippel. Building a historical corpus for Classical Portuguese: Some technological aspects. In: Proceedings of LREC. (2006) 1831-1836

Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK. (1994) 44-49

Tang, K., Nevins, A. Quantifying the diachronic productivity of irregular verbal patterns in Romance. Vol. 25. UCL Working Papers in Linguistics. (2013) 289-308

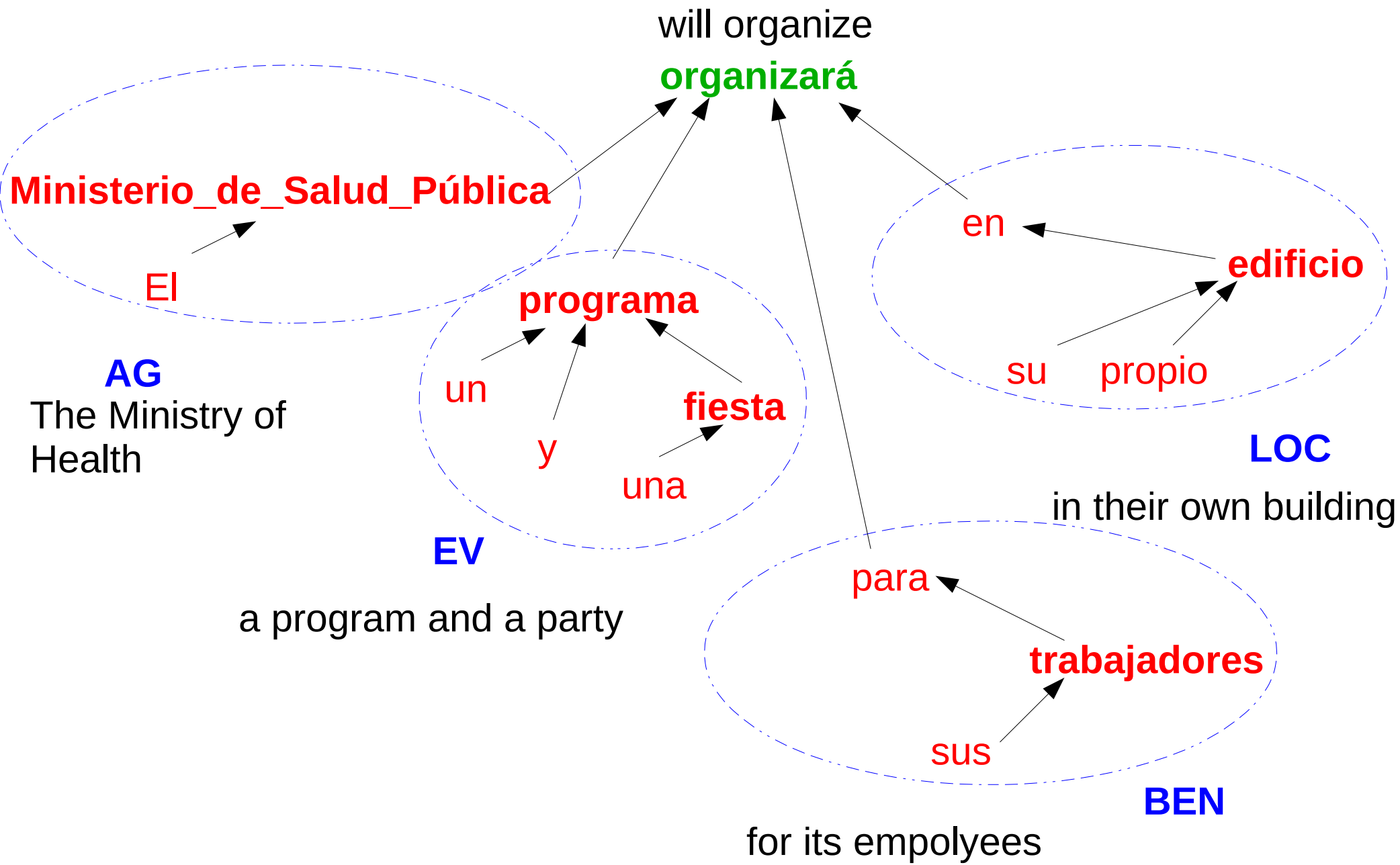
Zampieri, M., Becker, M. Colonia: Corpus of Historical Portuguese. In: ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research. Volume 5. Shaker. (2013) 69-76

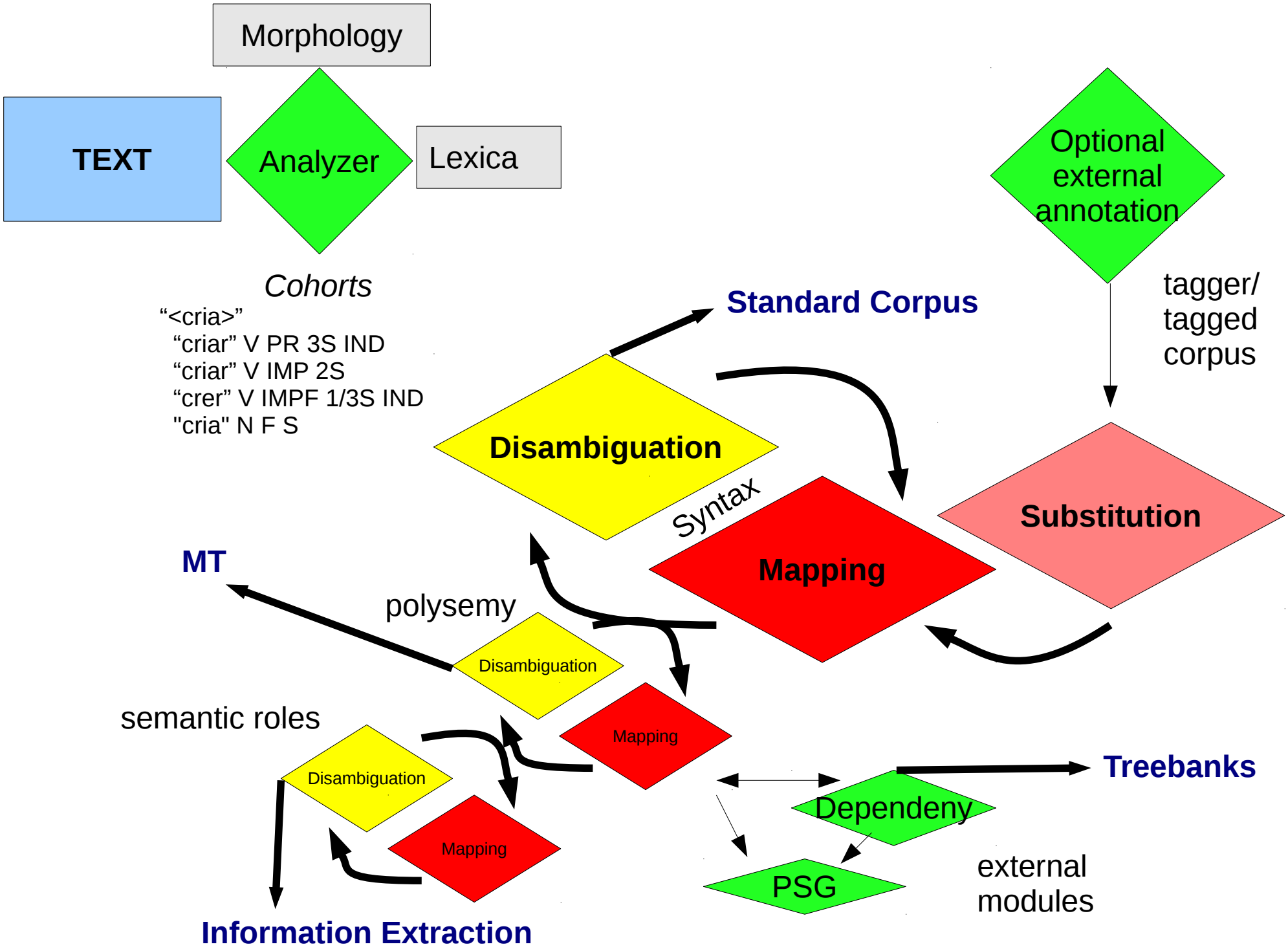
The point of departure: PALAVRAS

- Modular Constraint Grammar (CG) parser with a hierarchically structured sets of contextual rules
  - morphosyntactic tagging, dependency trees
  - 6000 rules, full lexical support, semantics

O	<artd>	DET M S	@>N	#1->3	*The*
último	<num-ord>	ADJ M S	@>N	#2->3	*last*
diagnóstico	<sem-c>	N M S	@SUBJ>	#3->9	*diagnostic*
elaborado	<pass>	V PCP2 M S	@ICL-N<	#4->3	*produced*
por		PRP	@<PASS	#5->4	*by*
a	<artd>	DET F S	@>N	#6->7	*the*
Comissão=Nacional	<org>	PROP F S	@P<	#7->5	*CN*
não	<neg>	ADV	@ADVL>	#8->9	*not*
deixa	<vt>	V PR 3S	@FMV	#9->0	*leaves*
dúvidas	<sem-c>	N F P	@<ACC	#10->9	*doubts*
$.				#11->0	

Dependency trees




Morphology
TEXT
Analyzer
Lexica
Optional external annotation
Cohorts
“<cria>”
“criar” V PR 3S IND
“criar” V IMP 2S
“crer” V IMPF 1/3S IND
"cria" N F S
Standard Corpus
tagger/ tagged corpus
Disambiguation
Syntax
Substitution
Mapping
MT
polysemy
Disambiguation
semantic roles
Mapping
Disambiguation
Treebanks
Dependeny
Mapping
PSG
external modules
Information Extraction

Text flow normalisation

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

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

Text flow: problems

- nesting and overlapping markers (the latter also problematic in xml)
- focus marker *é_que (2% of turns)* transcribed as *que* -> need for disambiguation
- syntax needs (separate) prepositions
  - built-in ordinary contractions: *do, nele, pelo ...*
  - corpus-specific: *pa (pra), pro, pum, naquea ...*
  - difficult ambiguity: *pra (para vs. para_a)*
- post-tokenization *(coral-inter)* with support from normalization lexicon for the most difficult contractions, e.g. *né = não é*

Lexical and orthographical normalization

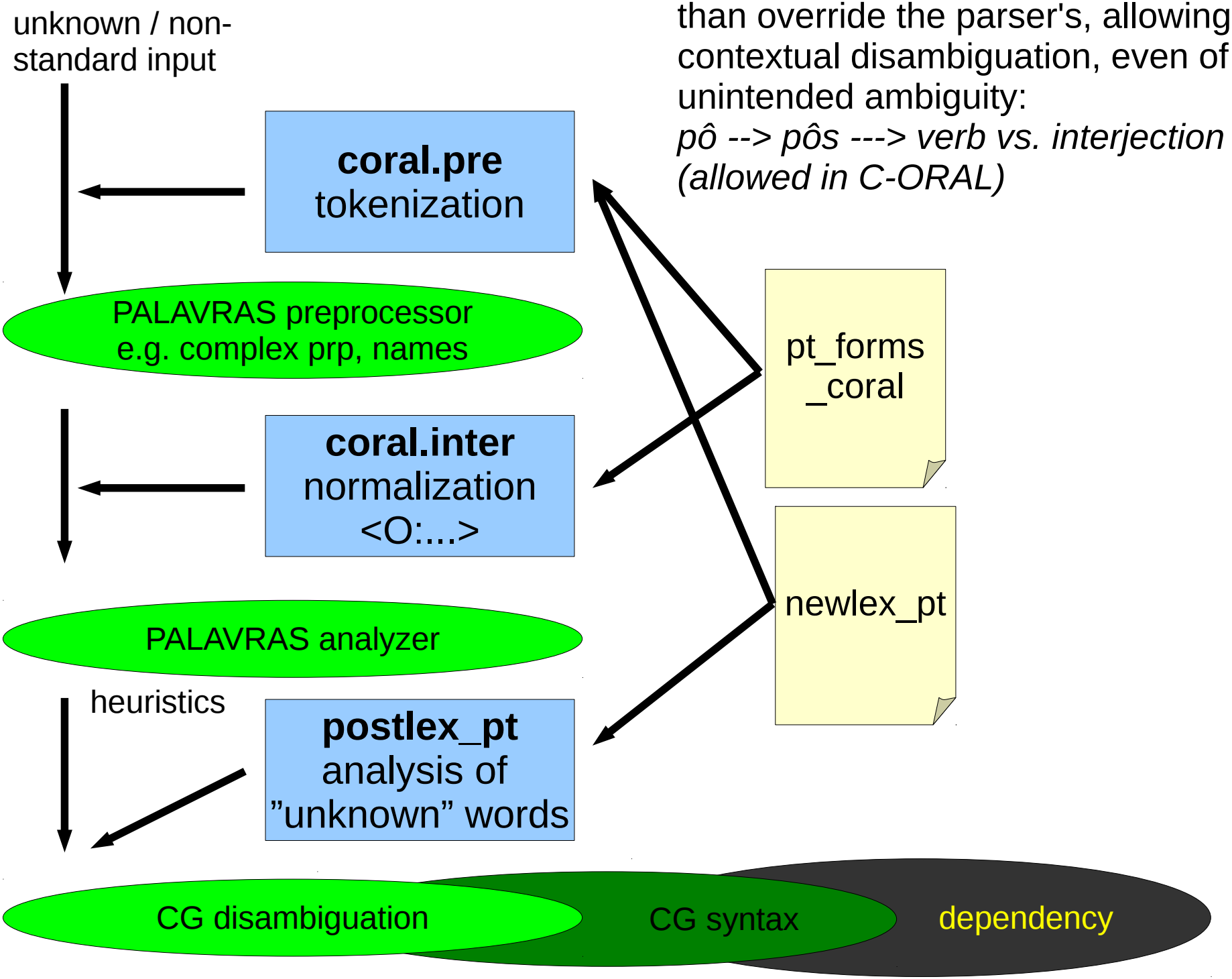
- Parser's treatment of unknown wordforms:
  - Affix-based derivations
  - Variants: br vs. pt, accents, orthographical reforms
- Special needs for C-ORAL speech corpus
  - "phonetically" transcribed word forms *(aquelas -> aqueas)*
  - grammatical variants *(-amos -> -amo)*
- Solutions
  - two-level annotation and specialized standardisation modules
    - *meninim OALT menininho [menino] N M S*
  - *coral.inter:* second preprocessor with systematic and item-based changes and MWE-tokenization
    - a'=qui (olha aqui), cabou (acabou)
  - *postlex_pt:* postprocessor with morphological analyzer using separate lexicon (2000 entries) and overriding PALAVRAS' heuristics

(a1) emedebê MDB (phonetic abbreviations)
(b3) inda ainda ((word-initial changes)
(b4) roz arroz
(d2) fazido feito (overregularization)

- Multi-word strings: effect also tokenization, but help disambiguate their parts, e.g.
  n' = *não (*not *em)* in: *n'=era, n'=ocê*
- Non-systematic new words and names:

(a1) fazeção <activity> N F S
(a2) zenes N M P # termo de jogo
(a3) caça-talentos N M S
(a5) superbem-arrumada ADJ F S
(b) mil-oitocentos-e=vovó=gostosa NUM M/F P
(c1) remote N M S # estrangeirismo
(c2) completed ADJ M/F S/P # estrangeirismo
(c5) anche ADV # estrangeirismo
(d1) tu=tu X # onomatopéia
(e2) TIM <org> PROP F S # company
(e3) Timoftol <cm-rem> PROP M S

- Add-on lexicon used to *add* readings rather than override the parser's, allowing for contextual disambiguation, even of unintended ambiguity:
  pô --> pôs ---> verb vs. interjection plural (allowed in C-ORAL)



Syntax

- Problem: syntactic noise: *ah, eeh, uh*
  - Solution: two-level annotation
- Problem:Syntactic annotation needs long-distance contexts, so how can existing rules be made to work on a speech corpus
  - > 80% unbounded/global CG rules in syntax
  - but the corpus lacks sentence segmentation and punctuation to delimit these rules
  - Solution: Exploit prosodic information by **not** moving it to a meta-level, but rather change it into punctuation
- // (major prosodic break) --> semicolon
- / (soft prosodic break) --> <break> <pause>

prosodic ”break markers”: rule-based disambiguation

- <break> --> comma
- <pause> --> meta-level
  - (a) between a noun or a nominative pronoun or a conjunction to the left, and a finite verb to the right, a prosodic /-marker is treated as <pause> (subject - verb case)
  - (b) prosodic /-markers between a noun and another np are treated as <break> (appositions)
  - (c) / between a prenominal and its head is treated as <pause> (np cohesion), e.g. 388 cases of article + <pause>

... <break> tipo **<retract:José>** Zé=Mourinho **<break>** falando assim **<break>** não **<break>**

o [o] <artd> DET M S @>N
<pause>
<campeonato> [campeonato] <occ> N M S @SUBJ>
d' OALT de [de] PRP @N<
ocês OALT vocês [você] PERS M/F 3P NOM/PIV @P<
é [ser] <vK> V PR 3S IND VFIN @FMV

Evaluation

- random transcription file (1895 tokens)
- eval_cg tool raw analysis file and revised version
- challenge: alignment in the face of punctuation ambiguity

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

Effectiveness of using prosodic break markers as punctuation

- standard run: pause/break disambiguation
- no-break: /-marks ignored
- no-sentence: both /- and //-marks ignored
- all-break: all /-marks as commas, no disambiguation

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

using prosodic breaks for syntax: results

- prosodic break markers do help the parser
- more so for syntax than PoS/morphology (wider contextual scope with corresponding segmentation needs)
- pause/break disambiguation more relevant for syntax than PoS

Exchange and export formats

C-ORAL dependency annotation in xml format

```
<sentence id=20>
        <word id="1" form="No'" base="Nossa" postag="intj" morf="--" extra="newlex" head="0" deprel="fA"/>
        <word id="2" form="," base="--" postag="pu" morf="--" extra="--" head="0" deprel="PU"/>
        <word id="3" form="o" base="o" postag="pron-indef" morf="--" extra="artd" head="4" deprel="DN"/>
        <word id="4" form="Galáticos" base="Galáticos" postag="prop" morf="--" extra="org newlex *" head="5" deprel="S"/>
        <word id="5" form="é" base="ser" postag="v-fin" morf="--" extra="-head vK fmc mv" head="1" deprel="CJT"/>
        <word id="6" form="mesmo" base="mesmo" postag="adv" morf="--" extra="quant" head="5" deprel="fA"/>
        <word id="7" form="," base="--" postag="pu" morf="--" extra="--" head="0" deprel="PU"/>
        <word id="8" form="todo_mundo" base="todo_mundo" postag="spec" morf="--" extra="--" head="9" deprel="S"/>
        <word id="9" form="é" base="ser" postag="v-fin" morf="--" extra="vK fmc mv" head="5" deprel="CJT"/>
        <word id="10" form="babaca" base="babaca" postag="n" morf="--" extra="Hfam" head="9" deprel="Cs"/>
        <word id="11" form=";" base="--" postag="pu" morf="--" extra="--" head="0" deprel="PU"/>

</speaker>
<speaker="GIL">
```

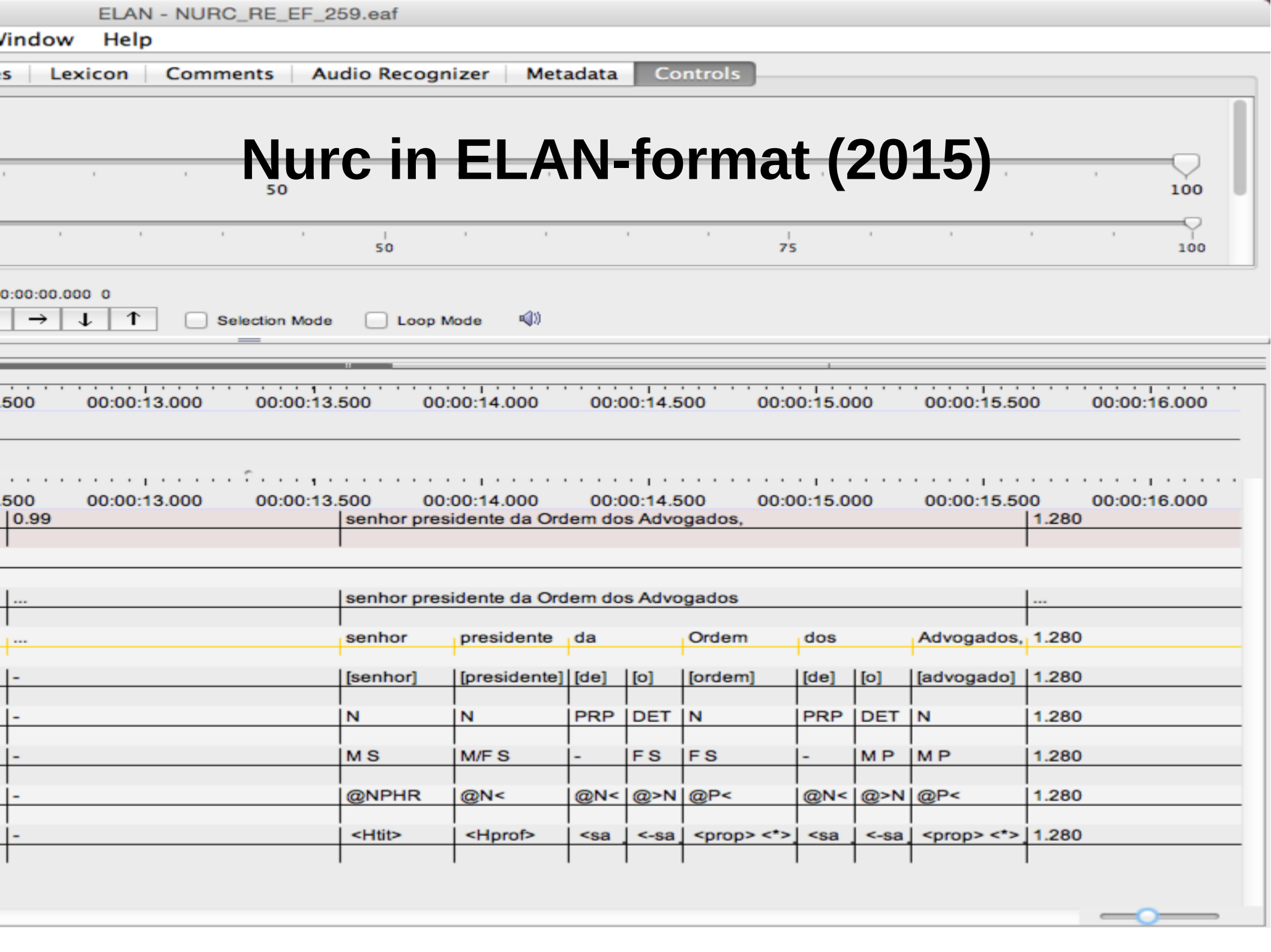
```
<inq INQ1> Porque é que... .  Desculpe, só... porque é que


<inf> INF1 </inf>


STA:fcl
=SUBJ:np
==>N:pron-det(<dem> M S)          Aquele
==H:n(M S)        buraco
=MV:v-fin(PR 3S IND)     enfia
=ADVL:pp
==H:prp(<sam->) em
==P<:np
===>N:art(<-sam> <artd> M S)      o
===H:n(M S)       tolete
=CO:conj-c        e
=FOC:v-fin        é_que
=ACC:adj(F S)     segura
=N<:np
==>N:art(<artd> M S)      o
==H:n(M S)        remo
==N<:adj(M S)     direito
=ADVL:prp         para
=P<:icl
==AUX:v-inf       poder
==MV:v-inf        remar
=.
```

Annotation alternatives:

VISL Constituent trees for CORDIAL-SIN:

exports to xml, PENN, TIGER, MALT, CQP ...

(Vila Praia de Âncora, 1999)


ELAN - NURC_RE_EF_259.eaf
Window Help
Lexicon Comments Audio Recognizer Metadata Controls
Nurc in ELAN-format (2015)
50
100
50
75
100
0:00:00.000 0
Selection Mode
Loop Mode
00:00:13.000
00:00:13.500
00:00:14.000
00:00:14.500
00:00:15.000
00:00:15.500
00:00:16.000
0.99
senhor presidente da Ordem dos Advogados,
1.280
...
senhor presidente da Ordem dos Advogados
...
senhor presidente da Ordem dos Advogados,
1.280
[senhor] [presidente] [de] [o] [ordem] [de] [o] [advogado]
1.280
N N PRP DET N PRP DET N
1.280
M S M/F S - F S F S - M P M P
1.280
@NPHR @N< @N< @>N @P< @N< @>N @P<
1.280
<Htit> <Hprof> <sa <-sa <prop> <*> <sa <-sa <prop> <*>
1.280

NURC: time-aligned xml & ELAN

cf. Oliveira & da Silva 2015
Projeto NURC Digital
IX LABLITA, Belo Horizonte

```
quero     [querer] <fmc> <vH> <mv> V PR 1S IND VFIN @FS-STA
<00:02:39.075 159.075 00:02:39.384 159.384>
me        [eu] <refl> PERS M/F 1S ACC @ACC>
<00:02:39.384 159.384 00:02:39.693 159.693>
referir   [referir] <vH> <mv> V INF @ICL-<ACC
$,
<00:02:39.694 159.694 00:02:40.881 160.881>
<length="1.187">
<00:02:40.881 160.881 00:02:41.133 161.133>
creio     [crer] <fmc> <vH> <mv> V PR 1S IND VFIN @FS-STA
<00:02:41.133 161.133 00:02:41.385 161.385>
que       [que] <clb-fs> KS @SUB
<00:02:41.385 161.385 00:02:41.637 161.637>
todos     [todo] <quant> DET M P @SUBJ>
<00:02:41.637 161.637 00:02:41.889 161.889>
já        [já] ADV @ADVL>
<00:02:41.889 161.889 00:02:42.141 162.141>
sabem     [saber] <mv> V PR 3P IND VFIN @FS-<ACC
<00:02:42.141 162.141 00:02:42.393 162.393>
de        [de] PRP @<PIV
<00:02:42.393 162.393 00:02:42.645 162.645>
```

```
00:00:11.268    11.268  00:00:12.590    12.59   excelentissimo  [excelente]      <SUP> <jh>      ADJ     M S     @PRED>
00:00:12.590    12.59   00:00:13.583    13.583  $break  -       -       -       -       -
00:00:13.583    13.583  00:00:13.925    13.925  senhor  [senhor]         <Htit> N        M S     @NPHR
00:00:13.925    13.925  00:00:14.267    14.267  presidente      [presidente]     <Hprof>         N       M/F S   @N<
00:00:14.267    14.267  00:00:14.438    14.438  da      [de]     <sam-> PRP      -       @N<
00:00:14.438    14.438  00:00:14.609    14.609          [o]      <-sam> <artd>  DET     F S     @>N
00:00:14.609    14.609  00:00:14.951    14.951  Ordem   [ordem]  <prop> <*> <act-s> <HH> <sit> <ac>     N       F S     @P<
00:00:14.951    14.951  00:00:15.122    15.122  dos     [de]     <sam-> PRP      -       @N<
00:00:15.122    15.122  00:00:15.293    15.293          [o]      <-sam> <artd>  DET     M P     @>N
00:00:15.293    15.293  00:00:15.635    15.635  Advogados       [advogado]       <prop> <*> <Hprof>      N       M P     @P<
00:00:15.635    15.635  00:00:15.635    15.635  $,      -       -       -       -       -
00:00:15.638    15.638  00:00:16.917    16.917  1.280   1.280   1.280   1.280   1.280   1.280
00:00:16.917    16.917  00:00:17.547    17.547  secção  [secção]         <ALT:seção> <inst> <Labs> <act> <geom> <sem>   N
```

Part 2: Speech-like Corpora

- Spoken language data are difficult to obtain in large quantities (very time & labour consuming)
- Hypothesis: Certain written data may approximate some of the linguistic features of spoken language
  - Candidates: chat, e-mail, broadcasts, speech and discussion transcripts, film subtitle files
- Topics
  - Suitable/available corpora
  - Tokenization and annotation methodology
  - linguistic insights and cross-corpus comparison

The corpora

- **Enron E-mail Dataset**: corporate e-mail (CALO Project)
- **Chat Corpus** 2002-2004 (Project JJ)
  - (a) Harry Potter, (b) Goth Chat, (c) X Underground, (d) Amarantus: War in New York
- **Europarl** - English section (Philipp Koehn)
  - transcribed parliamentary debates
- **BNC** (British National Corpus)
  - split in (a) written and (b) spoken sections

Annotation: Constraint Grammer - EngGram (CG3)
(demo: http://visl.sdu.dk/en/)

CG adaptations for speech-like data

- even a robust parser will suffer a performance decrease when ported from written to data with oral language traits
- CG does not need hand-corrected training corpora (which would be hard to find cross-domain, or with unified tagset)
- CG guarantees complete cross-domain compatibility, while at the same time allowing specific and repeated domain adaptations
  - Imperatives --> context rules & lexical statistics
  - Questions --> context rules
  - oral genre-specific items: interjections, emoticons (smileys)
    --> lexicon additions (e.g. *grg, oy*)
    --> heuristics for "productive" interjections (e.g. *oh ooh oooh, uh uh-uh*)
  - 1. and 2. pronoun frequency, "I"-disambiguation

Imperative vs. infinitive and present tense

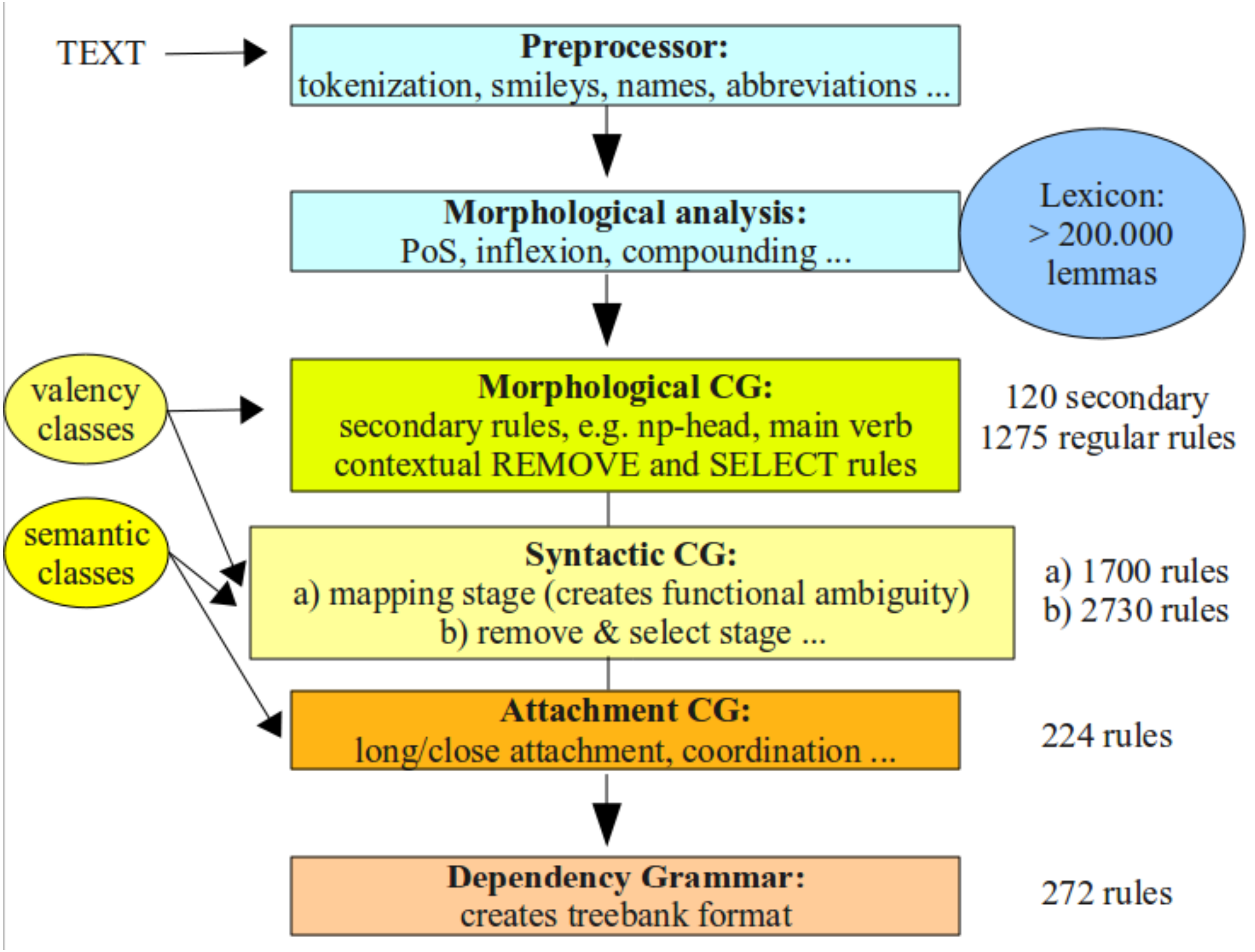
- written language parsers have an anti-imperative bias
- use context to disambiguate imperatives more precicely

```
SELECT (IMP) IF
  (-1 KOMMA) (*-2 VFIN BARRIER CLB
  LINK *-1 ("if") BARRIER CLB OR VV LINK *-1 >>> BARRIER NON-ADV/KC)
```

- use lexical likelihood statistics from mixed corpora
  - "<add>"
    - "add" <fr:12> V IMP
    - "add" <fr:68> V PR -3S
    - "add" <fr:20> V INF
  - "<achieve>"
    - "achieve" <fr:0> V IMP
    - "achieve" <fr:4> V PR -3S
    - "achieve" <fr:96> V INF

Parsing architecture

- multiple modularity
  - emoticon etc. preprocessing + morphological analysis + CG
  - multi-stage CG with rule sets at progressive levels with different annotation tasks
  - within each level: rule batches with increasing heuristicity, i.e. safe rules first: 1-2 ... 1-2-3 ... 1-2-3-4 ... 1-2-3-4-5 etc.
- lexicon support at all levels, both pos and syntax
  - valency: <vt>, <+on>, <+INF>, <vtk+ADJ>
  - semantic prototypes for nouns <Hprof>, <tool> and some adjectives <jnat> (nationhood), <jgeo> (geographical)
- highest level in this project is a kind of live dependency treebank, with all words linked to other words


TEXT
Preprocessor:
tokenization, smileys, names, abbreviations ...
Morphological analysis:
PoS, inflexion, compounding ...
Lexicon:
> 200.000
lemmas
valency
classes
Morphological CG:
secondary rules, e.g. np-head, main verb
contextual REMOVE and SELECT rules
120 secondary
1275 regular rules
semantic
classes
Syntactic CG:
a) mapping stage (creates functional ambiguity)
b) remove & select stage ...
a) 1700 rules
b) 2730 rules
Attachment CG:
long/close attachment, coordination ...
224 rules
Dependency Grammar:
creates treebank format
272 rules

Cross-corpus parser evaluation

- pilot evaluation with small data sets
- "soft" gold standard, created from parser output rather than from scratch, no multi-annotator cross-evaluation

	Chat 921			Enron e-mail 1078 tokens			Europarl 1446 tokens		
	R	P	**F**	R	P	**F**	R	P	**F**
PoS	93.2	93.2	**93.2**	98.3	98.3	**98.3**	99.7	99.7	**99.7**
syntactic function	87.5	88.5	**87.9**	93.3	92.5	**92.8**	95.2	96.6	**95.8**

Problems with oral-specific traits (especially chat corpus)

- Contractions:
  - *dont, gotta*
- "phonetic" writing:
  - *Ravvvvvvvveeee*
- unknown or drawn-out interjections read as nouns:
  - *tralalalala*
- unknown non-noun abbreviations
  - *sup (adjective), rp (infinitive), lol (interjection)*
- Subject-less sentences
  - *dances about wild and naked* ('dances' misread as noun)

Cross-corpus comparison of orality markers

- because CG annotation is token based at all levels, even higher-level syntactic information can be used
- BNC-written included as a kind of reference corpus for the orally-influenced text types
- expected differences along a "linguistic complexity" axis:
  - **chat < e-mail < Europarl < BNC-oral < BNC-written**
- high-complexity markers:
  - verb chain length, sentence length, subordination / subclauses, would/should-distancing, passive/active ratio for participles
- low-complexity markers:
  - interjections, pronouns

	Chat	E-mail	Euro-parl	BNC spoken	BNC written
function words	20.0 M	82.5 M	24.8 M	18.9 M	48.1 M
av. sentence length	8.74	19.71	**21.61**	17.27	18.12
av. word length	**4.4**	5.07	**5.27**	4.92	4.97
finite subclauses	4.32	**3.28**	4.29	**4.43**	4.09
relative	**1.96**	1.72	1.84	1.65	**1.57**
accusative	0.78	**0.64**	1.12	**1.28**	1.01
adverbial	1.25	**0.63**	0.93	**1.18**	1.12
gerund subclauses	**2.61**	1.43	**1.1**	1.2	1.3
infinitive subclauses	**1.57**	2.45	**2.48**	1.86	1.86
past part. subclauses	**0.21**	**0.42**	**0.37**	**0.21**	**0.22**
auxiliaries	**2.71**	5.06	**5.13**	4.10	3.79
active pcp2	**0.27**	0.55	0.72	**0.79**	**0.76**
passive pcp2	**0.33**	1.28	**1.48**	1.26	1.22
coordinating conj.	**3.14**	**3.36**	**3.52**	**3.56**	**3.76**
subordinating conj.	**1.33**	1.65	**2.04**	1.81	1.6
vocative	0.01	0	0.01	0.01	0.01
imperative	**0.35**	**0.5**	**0.05**	0.27	0.28
would, should, could	**0.41**	0.64	**0.8**	0.54	0.49
interjections	**0.92**	0.03	**0.01**	**0.56**	**0.1**
demonstrative	**1.04**	1.36	**2.23**	1.21	1.06
attributive	**5.15**	**5.51**	**7.51**	**7.74**	**8.42**
common nouns	25.61	**28.54**	**20.81**	21.71	22.62
proper nouns	**2.28**	**2.25**	3.89	4.18	**4.76**
finite verbs	**10.48**	10.21	9.36	10.92	10.47
personal & possessive pronouns	**12.36**	**3.32**	5.55	7.06	5.86

- **Chat** data is most consistently oral
- **Europarl/Enron** > BNC for aux, passive pcp and would/should
  --> complex oral style
- **Europarl** =monologue
  - longest w and s
  - subordination
  - inf / pcp - clauses
- **BNC** oral ~ written
  - only small differ.
  - high active pcp
    --> narrative
    adj and prop
    --> descriptive