Annotating Corpora for Linguistics
from text to knowledge

Eckhard Bick

University of Southern Denmark

Research advantages of using a corpus rather than introspection

- **empirical, reproducable:** Falsifiable science
- **objective, neutral:** The corpus is always (mostly) right, no interference from test-person's respect for textbooks
- **definable observation space:** Diachronics, genre, text type
- **statistics:** Observe linguistic tendencies (%) as opposed to (speaker-dependent) “stable” systems, quantify ?, ??, *, **
- **context:** All cases count, no “blind spots”

Teaching advantages of using a corpus rather than a textbook

- **Greater variety of material, easy to find many comparable examples:** A teacher's tool
- **An instant learner's dictionary:** on-the-fly information on phrasal verbs, prepositional valency, polysemy, spelling variants etc.
- **Explorative language learning:** real life text og speech, implicit rule building, learner hypothesis testing
- **Contrastive issues:** context/genre-dependent statistics, bilingual corpora

How to enrich a corpus

- Meta-information, mark-up: Source, time-stamp etc.
- Grammatical annotation:
  - Part of speech (PoS) and inflexion
  - syntactic function and syntactic structure
  - semantics, pragmatics, discourse relations
- Machine accessibility, format enrichment, e.g. xml
- User accessibility: graphical interfaces, e.g. CorpusEye, Linguateca, Glossa

The contribution of NLP to corpus linguistics

- in order to extract safe linguistic knowledge from a corpus, you need
  - (a) as much data as possible
  - (b) search & statistics access to linguistic information, both categorial and structural
- (a) and (b) are in conflict with each other, because enriching a large corpus with markup is costly if done manually
- tools for automatic annotation will help, if they are sufficiently robust and accurate

corpus sizes

manual

- ca. 1-10K - teaching treebanks (VISL), revised parallel treebanks (e.g. Sofie treebank)
- ca. 10-100K - subcorpora in speech or dialect corpora (e.g. CORDIAL-SIN), test suites (frasesPP, frasesPB)
- ca. 100K - 1M: monolingual research treebanks (revised), e.g. CoNLL, Negra
  Floresta Sintá(c)tica
- ca. 1-10M - specialized text corpora (e.g. ANCIB email corpus, topic journal corpora, e.g. Avante!), small local newspapers (e.g. Diário de Coimbra)

- ca. 10-100M - balanced text corpora (BNC, Korpus90)
  most newspaper corpora (Folha de São Paulo, Korpus2000, Information),
  genre-corpora (Europarl, Rumanian business corpus, chat corpus, Enron e-mail)
- ca. 100M -1G - wikipedia corpora, large newspaper corpora (e.g. Público), cross-language corpora (e.g. Leipzig corpora)
- > 1G - internet corpora

automatic

corpus size and case frames (Japanese)

Table 4: Statistics of the Constructed Case Frames.

Corpus size (sentences)		1.6M	6.3M	25M	100M	400M	1.6G
# of predicate		2460	6134	13532	27226	42739	65679
(type)	verb	2039	4895	10183	19191	28523	41732
	adjective	154	326	617	1120	1641	2318
	noun with copula	267	913	2732	6915	12575	21629
average # of case frames for a predicate		15.9	12.2	13.3	16.1	20.5	25.3
average # of case slots for a case frame		2.95	3.44	3.88	4.21	4.69	5.08
average # of examples for a case slot		4.89	10.2	19.5	34.0	67.2	137.6
average # of unique examples for a case slot		1.19	1.85	3.06	4.42	6.81	9.64
average # of generalized examples for a case slot		0.14	0.24	0.37	0.49	0.67	0.84
File size(byte)		8.9M	20M	56M	147M	369M	928M

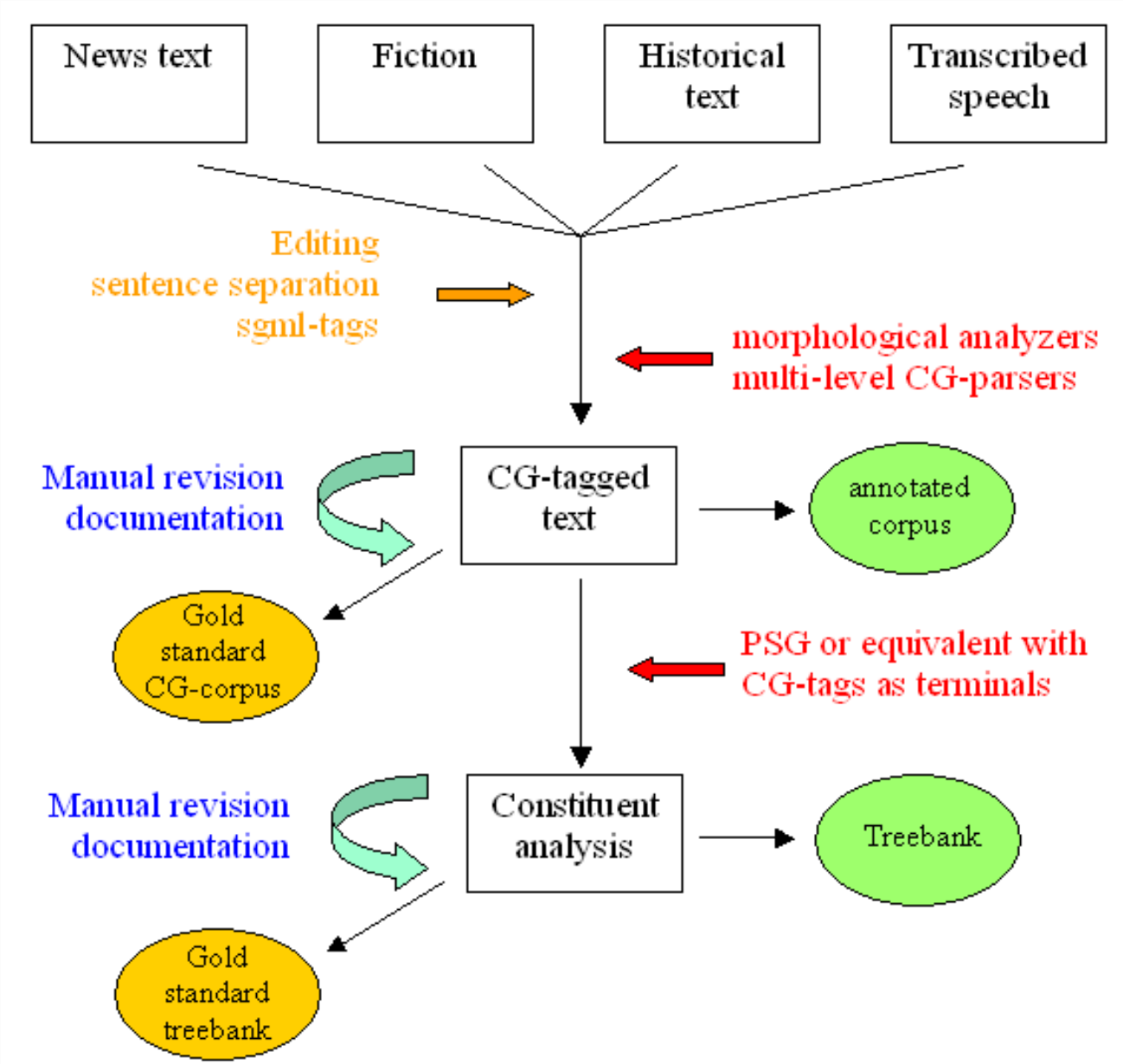
Sasano, Kawahara & Kurohashi: "The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis", in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*

The number of unique examples for a case slot increases ~ 50% for each fourfold increase in corpus size

Added corpus value in two steps, a concrete example:

1. annotation

2. revision



The neutrality catch

- All annotation is theory dependent, but some schemes less so than others. The higher the annotation level, the more theory dependent
- The risk is that "annotation linguistics" influences or limits corpus linguistics, i.e. what you (can) conclude from corpus data
- "circular" role of corpora: (a) as research data, (b) as gold-standard annotated data for machine learning: rule-based systems used for boot-strapping, will thus influence even statistical systems
- PoS (tagging): needs a lexicon (“real” or corpus-based)
  (a) probabilistic: HMM-base line, DTT, TnT, Brill etc., F-score ca. 97+%
  (b) rule-based:
  --- Disambiguation as a “side-effect” of syntax (PSG etc.)
  --- Disambiguation as primary method (CG), F-score ca. 99%
- Syntax (parsing): function focus vs. form focus
  (a) probabilistic: PCFG (constituent),
  MALT-parser (dependency F 90% after PoS)
  (b) rule-based: HPSG, LFG (constituent trees),
  CG (syn. function F 96%, shallow dependency)

Parsing paradigms:

Descriptive versus methodological (more "neutral"?)

Descriptive
Motivation: Explanatory
Test: teaching

Methodological
Robust
Machine translation

- **Generative** rewriting parsers: function expressed through structure
- **Statistical** taggers: function as a token classification task
- **Topological** “field” grammars: function expressed through topological form
- **Dependency** grammar: function expressed as word relations
- **Constraint Grammar**: function through progressive disambiguation of morphosyntactic context

Constraint Grammar

- A methodological parsing paradigm (Karlsson 1990, 1995), with descriptive conventions strongly influenced by dependency grammar
- Token-based assignment and contextual disambiguation of tag-encoded grammatical information, “reductionist” rather than generative
- Grammars need lexicon/analyzer-based input and consist of thousands of MAP, SUBSTITUTE, REMOVE, SELECT, APPEND, MOVE ... rules, that can be conceptualized as high level string operations.
- A formal language to express contextual grammars
- A number of specific compiler implementations to support different dialects of this formal language:

cg-1	Lingsoft 1995
cg-2	Pasi Tapainen, Helsinki University, 1996
FDG	Connexor, 2000
vislcg	SDU/Grammarsoft, 2001
vislcg3	Grammarsoft/SDU, 2006... (frequent additions and changes)

Differences between CG systems

- Differences in expressive power
  - scope: global context (standard, most systems) vs. local context (Lager's templates, Padró's local rules, Freeling ...)
  - templates, implicit vs. explicit barriers, sets in targets or not, replace (cg2: reading lines) vs. substitute (vislcg: individual tags)
  - topological vs. relational
- Differences of applicational focus
  - focus on disambiguation: classical morphological CG
  - focus on selection: e.g valency instantiation
  - focus on mapping: e.g. grammar checkers, dependency relations
  - focus on substitutions: e.g. morphological feature propagation, correction of probabilistic modules

The CG3 project

- 3+ year project (University of Southern Denmark & GrammarSoft)
- some external or indirect funding (Nordic Council of Ministries, ESF) or external contributions (e.g. Apertium)
- programmer: Tino Didriksen
- design: Eckhard Bick (+ user wish list, PaNoLa, ...)
- open source, but can compile "non-open", commercial binary grammars (e.g. OrdRet)
- goals: implement a wishlist of features accumulated over the years, and do so in an open source environment
- support for specific tasks: MT, spell checking, anaphora ...

Hybridisation: incorporating other methods:

- Toplogical method: native:
  - ±n position, * global offset, LINK adjacency, BARRIER ...
- Generative (rewriting) method: “Template tokens”
  - TEMPLATE np = (ART, ADJ, N) OR (np LINK 1 pp + @N<)
  - Feature/attribute Unification: $$NUMBER, $$GENDER ...
- Dependency:
  - SETPARENT (dependent_function) TO (*1 head_form) IF
- Probabilistic:
  - <frequency> tags, e.g. <fr:49> matched by <fr>30>

The CG3 project -2

- working version downloadable at http://beta.visl.sdu.dk
- compiles on linux, windows, mac
- speed: equals vislcg in spite of the new complex features, faster for mapping rules, but still considerably slower than Tapanainen's cg2 (working on it).
- documentation available online
- sandbox for designing small grammars on top of existing parsers: The cg lab

What is CG used for?

Lingsoft
LANGUAGE SOLUTIONS

Lingsoft Proofreader Finnish for Sisulizer 2008
99,00 € + VAT 23 %
Add to cart >
More Sisulizer versions...

Lingsoft Svefix 2.2 for Microsoft Office
50,00 € + VAT 23 %
Evaluate Add to cart >

Lingsoft Speech Controller 1.4 (Windows)
220,49 € + VAT 23 %
Evaluate Add to cart >

VISL grammar games:

connexor Natural Knowledge

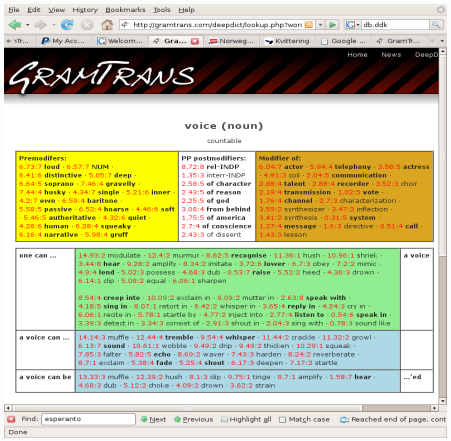
Machinese parsers
News feed and relevance filtering
Opinion mining in blogs
Science publication monitoring

GRAMTRANS

OrdRet

QA

Machine translation
Spell- and Grammar checking
Corpus annotation
Relational dictionaries: DeepDict



NER

Annotated corpora: CorpusEye

CG languages (VISL/GS)

	Parser	Lexicon	Grammar	Appl.	Corpora
	DanGram	100.000 lexemes, 40.000 names	8.400 rules	MT, grammar checker, NER, teaching, QA	ca. 150 mill. words (mixed, news)
	PALAVRAS	70.000 lexemes, 15.000 names	7.500 rules	teaching, NER, QA, MT	ca. 380 mill. words (news, wiki, europarl ...)
	HISPAL	73.000 lexemes	4.900 rules	teaching	ca. 86 mill. words (Wiki, Europarl, Internet)
	EngGram	81.000 val/sem	4.500 rules	teaching, MT	ca. 210 mill. words (mixed) 106 mill. email & chat
	SweGram	65.000 val/sem	8.400 rules	teaching, MT	ca. 60 mill. words (news, Europarl, wiki)
	NorGram	OBT / via DanGram	OBT / via DanGram	teaching, MT	ca. 30 mill. words (Wikipedia)
	FrAG	57.000 lexemes	1.400 rules	teaching	67 mill (Wiki, Europarl)
	GerGram	25.000 val/sem	ca. 2000 rules	teaching, MT	ca. 44 mill. words (Wiki, Europarl, mixed)
	EspGram	30.000 lexemes	2.600 rules	grammar checker, MT	ca. 40 mill. words (mixed, literature, internet, news)
	ItaGram	30.600 lexemes	1.600 rules	teaching	46 mill. (Wiki, Europarl)

VISL languages (others)

- Basque
- Catalan
- English ENGCG (CG-1, CG-2, FDG)
- Estonian (local)
- Finnish (CG-1?)
- Irish (Vislcg)
- Norwegian (CG-1, CG-3)
- Sami (CG-3)
- Swedish (CG1, CG-2?)
- Swahili (Vislcg)

Apertium “incubator” CGs

(https://apertium.svn.sourceforge.net/svnroot/apertium/...)

- **Turkish**: .../incubator/apertium-tr-az/apertium-tr-az.tr-az.rlx
- **Serbo-Croatian**: .../incubator/apertium-sh-mk/apertium-sh-mk.sh-mk.rlx
- **Icelandic**: .../trunk/apertium-is-en/apertium-is-en.is-en.rlx
- **Breton**: .../trunk/apertium-br-fr/apertium-br-fr.br-fr.rlx
- **Welsh**: .../trunk/apertium-cy-en/apertium-cy-en.cy-en.rlx
- **Macedonian**: .../trunk/apertium-mk-bg/apertium-mk-bg.mk-bg.rlx
- **Russian**: .../incubator/apertium-kv-ru/apertium-kv-ru.ru-kv.rlx

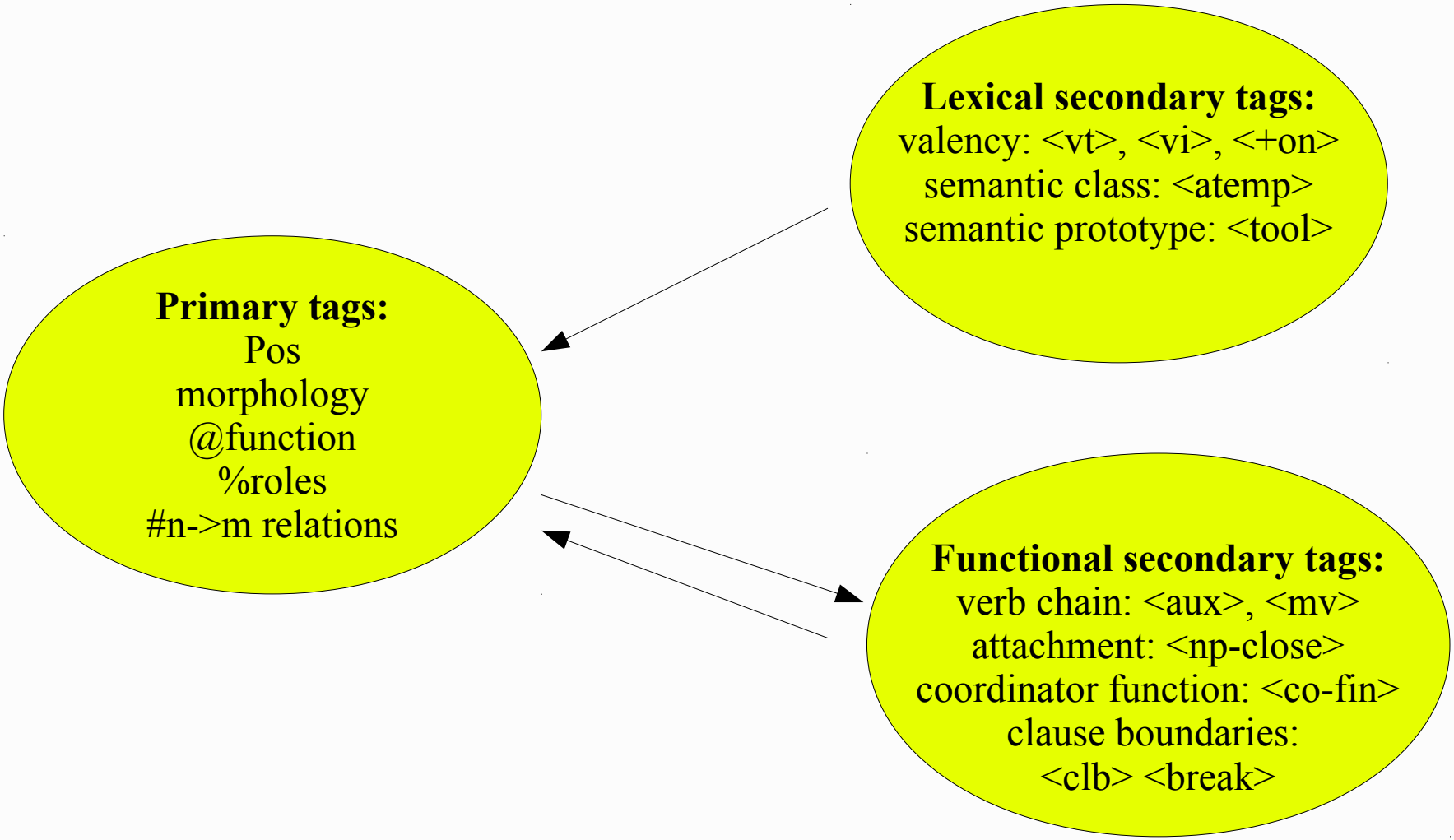
An output example: Numbered dependency trees

```
The        "the" <def> ART          @>N      #1->3
last       "last" ADJ               @>N      #2->3
report     "report" <sem-r> N S     @SUBJ>   #3->9
published  "publish" <vt> V PCP2    @ICL-N<  #4->3
by         "by" PRP                 @<PASS   #5->4
the        "the" <def>  ART         @>N      #6->7
IMF        "IMF" <org> PROP F S     @P<      #7->5
never      "never" ADV              @ADVL>   #8->9
convinced  "convince" <vt> V IMPF   @FMV     #9->0
investors  "investor" N F P         @<ACC    #10->9
$.                                           #11->0
```

Annotation principles - general

- **token-based tags**, also for structural annotation
- **discrete** rather than compound tags (e.g. CLAWS)
  - V PR 3S. not V-PR-3S or V3S
- **form & function** dualism at all levels
  - ADJ can function as np head without necessarily changing PoS category
  - pronoun classes are defined using inflexional criteria
  - syntactic function is independent of form, and established prior to bracketing or dependency (cp. labelled edges or chunk labeling strategies)
  - words have stable semantic (form) types, while being able to assume different semantic (function) roles

primary vs. secondary tags



Annotation - PoS and morphology

- **N (noun):** M,F,UTR,NEU - S,P - DEF,IDF - NOM,ACC,DAT,GEN...
- **ADJ (adjective):** = N + POS,COM,SUP
- **DET (determiner):** = N + <quant> <rel> <interr> <dem> ...
- **V (verb):** PR,IMPF,PS,FUT... - 123S,P - AKT,PAS - IND,SUBJ,IMP
  - INF, PCP1, PCP2 AKT, PCP2 PAS, PCP2 STA (=ADJ)
- **ADV (adverb):** COM,SUP
- **PERS (personal pronoun):** =N + 123S,P
- **INDP (independent pronoun):** S,P - NOM,ACC,..
- **other non-inflecting**: ART, NUM, PRP, KS, KC, IN

- Syntactic function annotation, clause level:
  - “case”-style function: @SUBJ, @ACC, @DAT
  - bound predicatives: @SC, @OC, @SA, @OA
  - free constituents: @ADVL, @PRED
  - meta constituents: @S<, @VOK, @TOP, @FOC
- group level:
  - np: @>N, @N<, @N<PRED, @APP
  - adjp, advp, detp: @>A, @A<
  - pp: @P<, @>P, @>>P, conjp: @>S
  - vp: @FMV, @IMV, @FAUX, @IAUX, @AUX<, @IMFM, @PRT, @MV<
- sub clause:
  - @FS- (finite), @ICL- (non-finite), @AS- (averbal)
- main clause: @STA, @QUE, @COM, @UTT

Annotation: structure

- shallow dependency
  - head-direction markers, e.g.: @SUBJ>, @<SUBJ, @>>P
  - secondary attament tags: <np-close>, <np-long>, <co-subj>, <co-fin>
- dependency trees
  - #n->m (n = ID daughter, m = ID head)
- constituent trees
  - clauseboundary markers: <clb> <cle>
  - vertical indentation notation, converted from dependency
- higher-level structure (arbitrary scope)
  - named relations x->y: ID=x REL:anaphor:y

Annotation: semantics

- semantic subclasses
  - adverbs: <atemp>, <aloc>, <adir>, <aquant> ....
  - pronouns: <rel>, <interr>, <dem>, <refl>, <quant> ...
- semantic prototypes
  - nouns: ~200 types: <Hprof>, <Vair>, <tool-shoot> ...
    - atomic feature bundles: ±hum, ±anim, ±move, ±loc ...
  - adjectives: <jnat> <jpsych> <jcol> <jshape> <jgeo> ...
- semantic roles
  - 15 core roles: §AG, §PAT, §TH, §REC, §COG ...
  - 35 “adverbial” and meta-roles: §DIR, §DES ....

CG rules

- rules add, remove or select morphological, syntactic, semantic or other readings
- rules use context conditions of arbitrary distance and complexity (i.e. other words and tags in the sentence)
- rules are applied in a deterministic and sequential way, so removed information can't be recovered (though I t can be traced). Robust because:
  - rules in batches, usually safe rules first
  - last remaining reading can't be removed
  - will assign readings even to very unconventional language input (“non-chomskyan”)

some simple rule examples

- REMOVE VFIN
  IF (*-1C VFIN BARRIER CLB OR KC)
  exploits the uniqueness principle: only one finite verb per clause

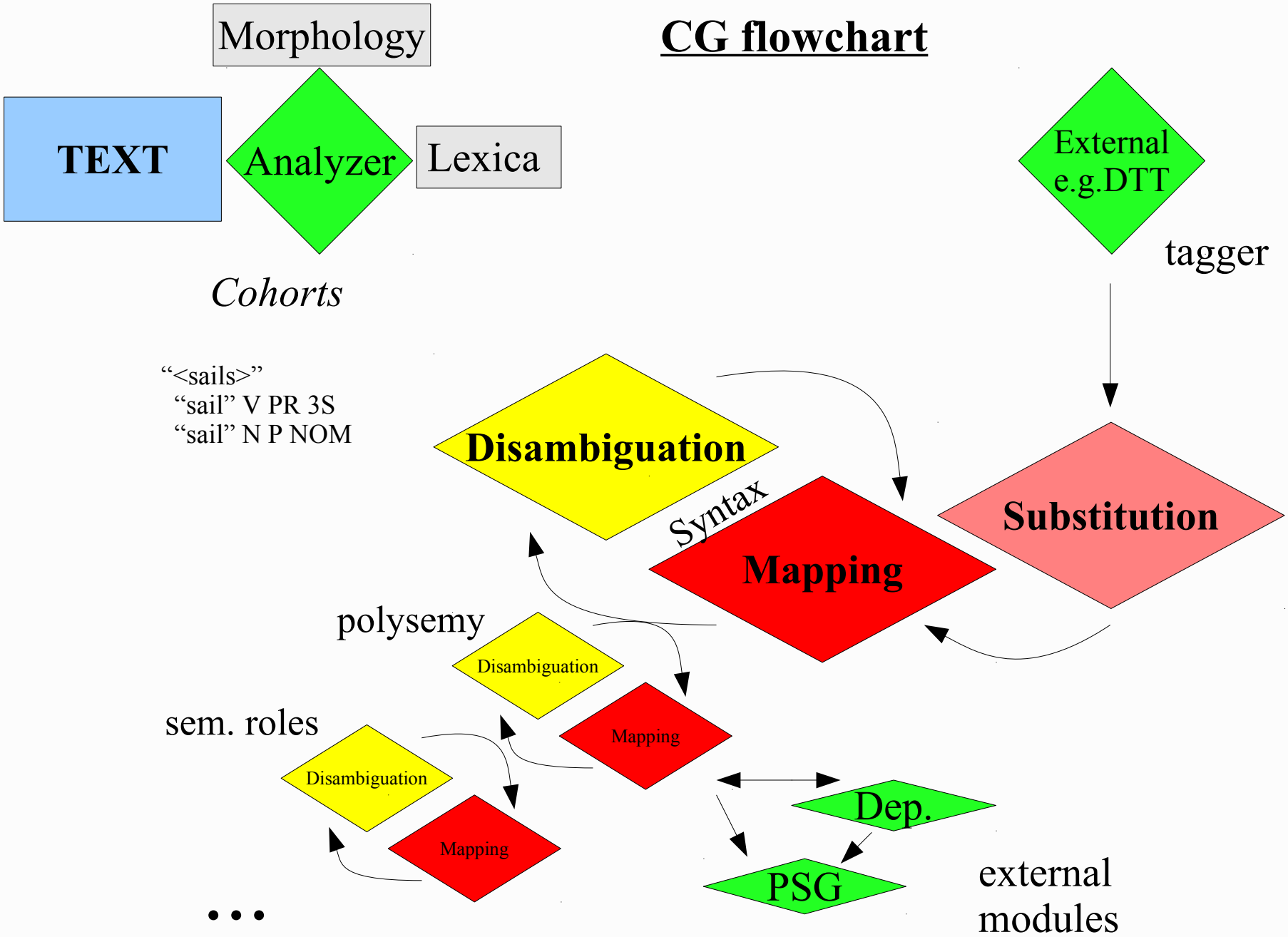
- MAP (@SUBJ> @<SUBJ @<SC) TARGET (PROP)
  IF (NOT -1 PRP)
  syntactic potential of proper nouns

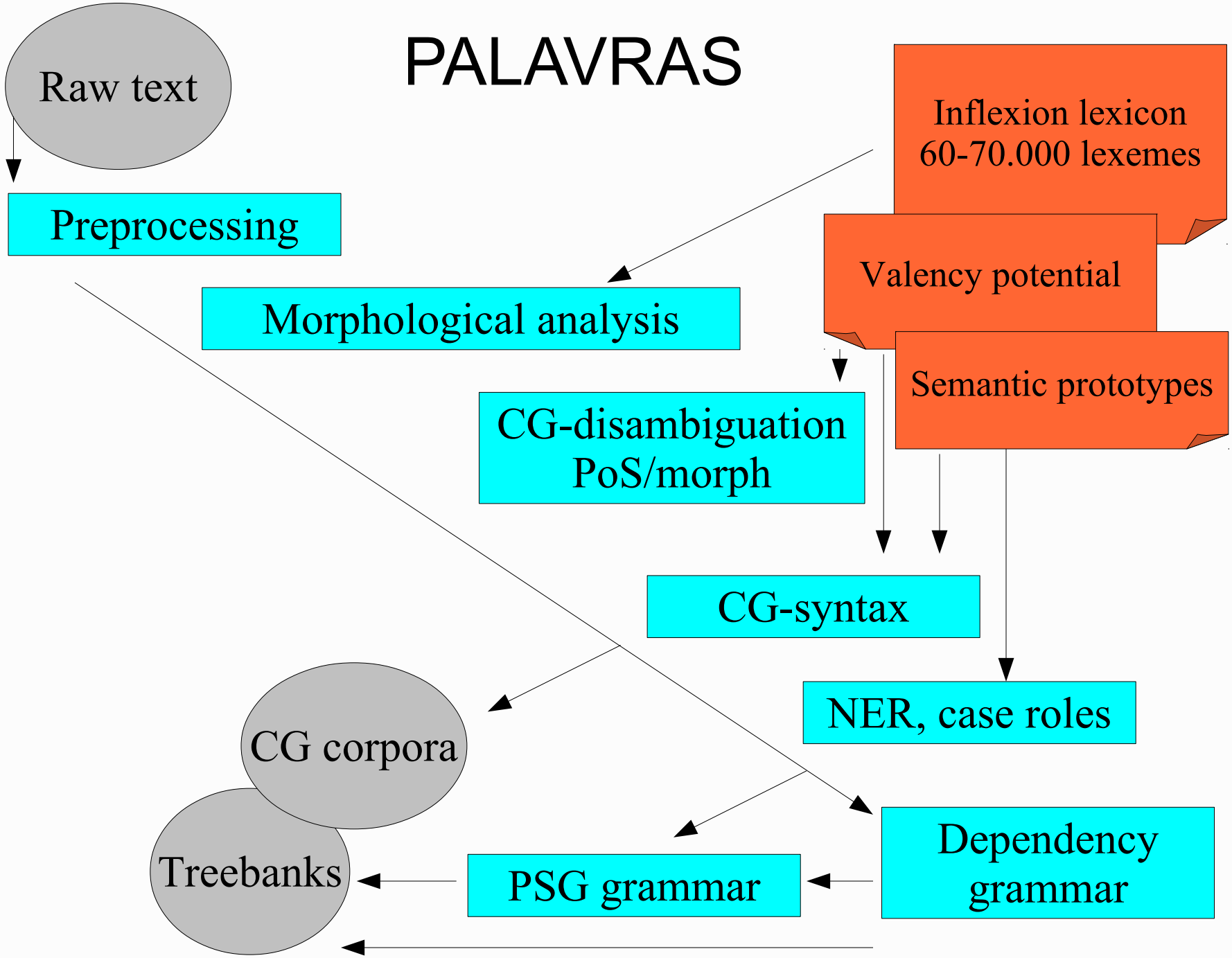
- SELECT (@SUBJ>)
  IF (*-1 >>> OR KS BARRIER NON-PRE-N/ADV)
  (*1 VFIN BARRIER NON-ATTR)
  clause-initial np's, followed by a finite verb, are likely to be subjects

- REMOVE (@<SUBJ)
  IF (NOT 0 N-HUM) (*-1 V-HUM BARRIER NON-PRE-N LINK 0 AKT) ;

- SELECT ADJ + MS
  IF (-1C ART + MS) (*2C NMS BARRIER NON-ATTR OR (F) OR (P)) ;

CG flowchart




PALAVRAS
Raw text
Preprocessing
Inflexion lexicon
60-70.000 lexemes
Morphological analysis
Valency potential
Semantic prototypes
CG-disambiguation
PoS/morph
CG-syntax
NER, case roles
CG corpora
Treebanks
PSG grammar
Dependency
grammar

The PALAVRAS system in current numbers

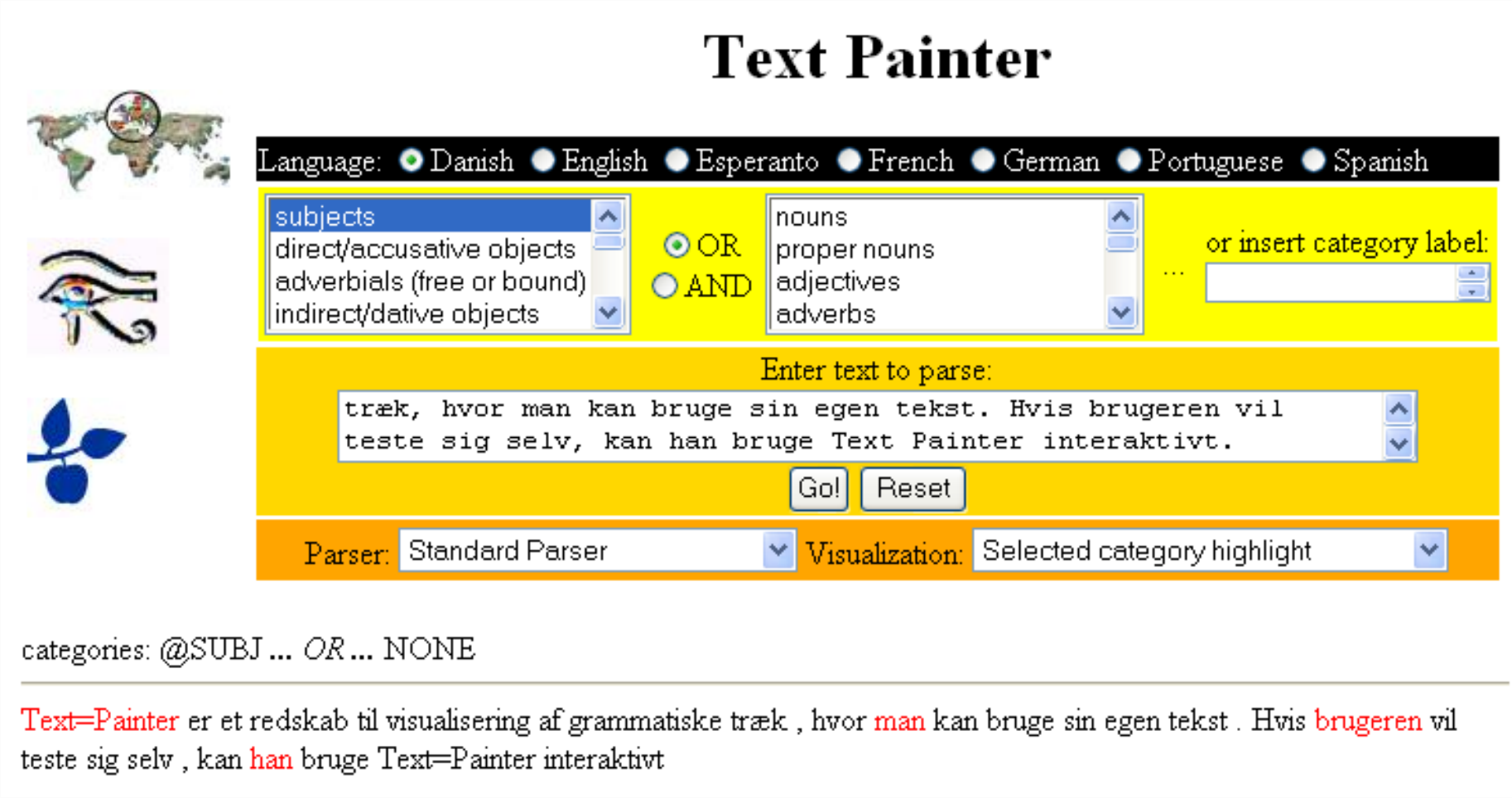
Lexemes in morphological base lexicon: ~70.000
(equals about 1.000.000 full forms), of these:
nouns with semantic prototypes: ~40.000
polylexicals: 9.000 (incl. some names)
Lexemes in the name lexicon: ~ 15.000
Lexemes in the frame lexicon: ~ 9.600 words

Portuguese CG rules, main grammar: 5.955
morphological CG disambiguation rules: 1.936
syntactic mapping-rules: 1.758
syntactic CG disambiguation rules: 2.261

Portuguese CG rules in add-on modules:4.921
valency instantiation rules and semantic type disambiguation: 3046
propagation rules: 614
attachment rules (tree structure preparing): 94
NER rules: 483
semantic roles: 397 (without dependency first: 514)
complex feature mapping ("procura" grammar): 75
Anaphora rules: 71
MT preparation rules (pt->da): 141
Portuguese PSG-rules: ~ 490 (for generating syntactic tree structures)
Portuguese Dependency-rules: ~ 260 (alternative way of generating syntactic tree structures)

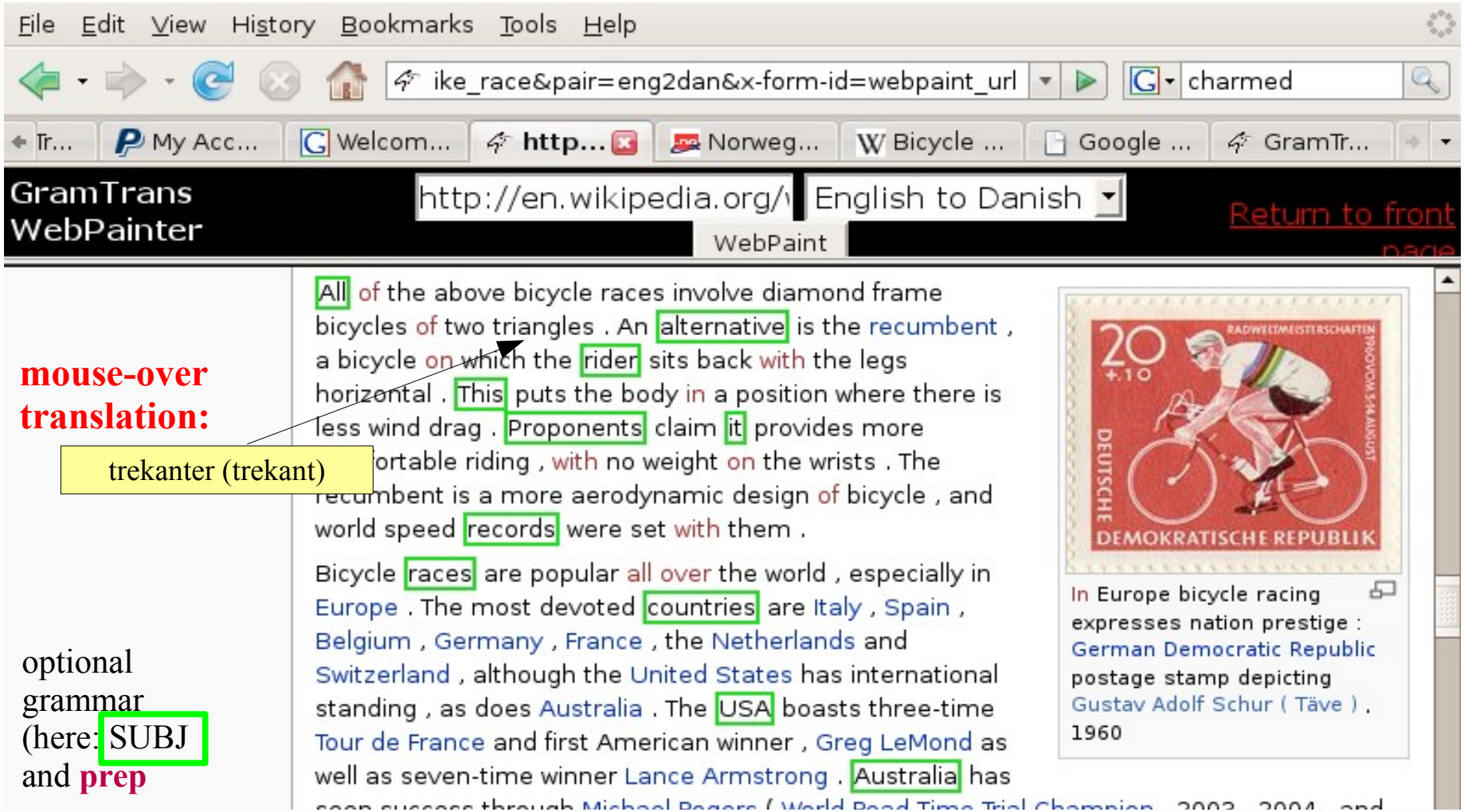
Performance:
At full disambiguation (i.e., maximal precision), the system has an average correctness of 99% for word class (PoS), and about 96% for syntactic tags (depending, on how fine grained an annotation scheme is used)
Speed:
full CG-parse: ca. 400 words/sec for larger texts (start up time a few seconds)
morphological analysis alone: ca. 1000 words/sec

Integrating live NLP and language awareness teaching

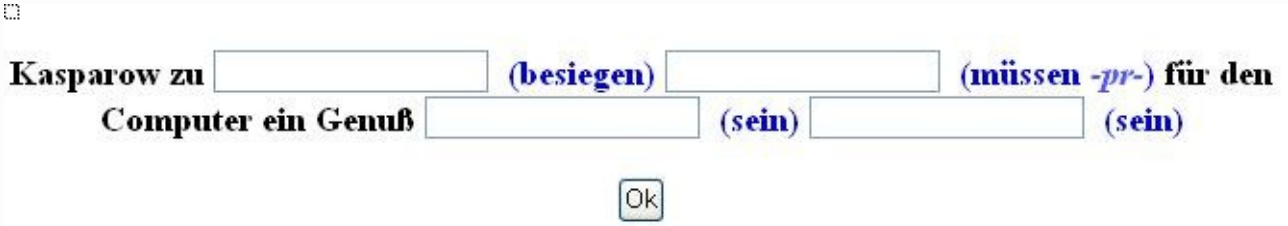


WebPainter

- live in-line markup of web pages
- mouse-over translations while reading



KillerFiller: Corpus-based, flexible slot-filler exercises



CG for corpus annotation

- can be used in modules, for raw text or for higher-level analysis on partially annotated corpora
- it normally needs morphological analysis as input, but can handle regular inflexion in the formalism itself
- speed for a big grammar, on a server-level computer is 15-20 million words / day
- since all information is expressed as word-based tags, it facilitates corpus query databases (CQP)

Annotated corpora (~1 billion words)

Annotated with morphological, syntactic and (some) dependency tags

- **Europarl**, parliament proceedings, 7 languages x 27M words (215M words)
- **Wikipedia**, 8 languages (~ 200M words)
- **ECI**, Spanish, German and French news texts, 14M words
- **Korpus90** and **Korpus2000**, mixed genre Danish, 56M words
- **Information,** Danish news text, ~ 80M words annotated
- **Göteborgsposten**, Swedish new text, ~ 60M words annotated
- **DFK**, mainly transscribed parliamentary discussions, 7M words
- **BNC,** balanced British English, 100M words
- **Enron**, e-mail corpus, 80M words
- **KEMPE**, Shakespeare historical corpus, 9M words
- **Chat**, English chat corpus, 24M words
- **CETEMPúblico**, European Portuguese, news text, 180M words
- **Folha de São Paulo**, Brazilian news text, 90M words
- **CORDIAL-SIN**, dialectal Portuguese, 30K words
- **NURC**, **C-ORAL-Brasil**, transscribed Brazilian speech, 100K words and 200K words
- **Tycho Brahe**, historical Portuguese, 50K words
- **RumBiz**, Rumanian business news, 9M words
- **Leipzig corpora**, mixed web corpora, various languages, ~20-30M each
- **Internet corpora**, Spanish (35M), Esperanto (28M)

Treebanks

- **Floresta Sintá(c)tica**, European Portuguese, 1M words (200K revised)
- **Arboretum**, Danish, 200-400K words revised
- **L'arboratoire**, French, ~ 20K words revised
- teaching treebanks for 25 languages (revised), 2K - 20K each
- unrevised "jungle" treebanks
  - **Floresta virgem**, 2 x 1M words Brazilian and European Portuguese
  - **Internet data treebanks,** various languages and sizes
  - **MT-smoother,** 1 billion words English mixed text

CG input

- Preprocessing
  - Tokenizer:
    - **Word-splitting**: *punctuation vs. abbreviation?, won't, he's vs. Peter's*
    - **Word-fusion:** *Abdul=bin=Hamad, instead=of*
  - Sentence separation: <s>...</s> markup vs. CG delimiters
- Morphological Analyzer
  - outputs cohorts of morphological reading lines
  - needs a lexicon and/or morphological rules

Integrating structure and lexicon: 2 different layers of semantic information

- (a) "lexical perspective": contextual selection of
  - a (noun) sense [**WordNet** style, http://mwnpt.di.fc.ul.pt/] or
  - semantic prototype [**SIMPLE** style, http://www.ub.edu/gilcub/SIMPLE/simple.html , http:/www.ub.es/gilcub/SIMPLW/simple.html]
- (b) "structural perspective": **thematic/semantic roles** reflecting the semantics of verb argument frames
  - Fillmore 1968: case roles
  - Jackendoff 1972: Government & Binding theta roles
  - Foley & van Valin 1984, Dowty 1987:
    - universal functors postulated
    - feature precedence postulated (+HUM, +CTR)

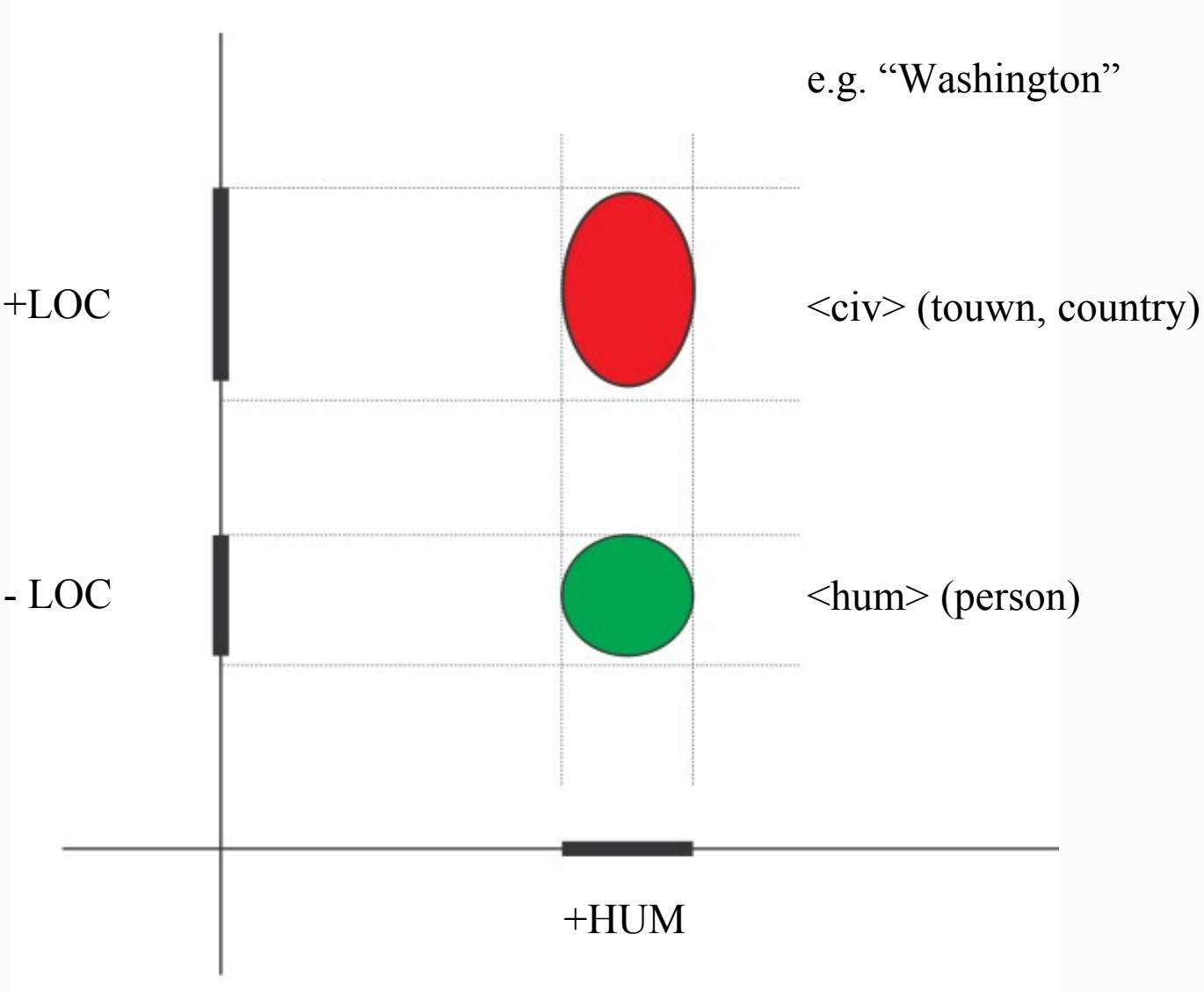
Semantic Annotation

- Semantic vs. syntactic annotation
  - semantic sentence structure, defined as a dependency tree of semantic roles, provides a more stable alternative to syntactic surface tags
- “Comprehension” of sentences
  - semantic role tags can help identify linguistically encoded information for applications like dialogue system, IR, IE and MT
- Less consensus on categories
  - The higher the level of annotation, the lower the consensus on categories. Thus, a semantic role set has to be defined carefully, providing well-defined category tests, and allowing the highest possible degree of filtering compatibility

what is a semantic prototype?

- semantic prototype classes perceived as distinctors rather than semantic definitions
- intended to at the same time
  - capture semantically motivated regularities and relations in syntax by similarity-lumping (syntax-restrictions, IR, anaphora)
  - distinguish different senses (polysemy)
  - select different translation equivalents in MT
- prototypes seen as (idealized) best instance of a given class of entities (Rosch 1978)
- *but*: class hypernym tags used (<Azo> for “land animal”) rather than low-level-prototypes (<dog> or <cat>)

Disambiguation of semantic prototype bubbles by dimensional downscaling (lower-dimension projections)



Semantic prototypes vs. Wordnet

- only ISA, no meronyms/holonyms/antonyms
- linguistic vs. encyclopaedic (dolphin, penguin)
- shallow vs. deep ontology, distinctional vs. definitiona

cavalo -- (Animals, Biology)
-> equídeos -- (Animals, Biology)
-> perissodáctilos -- (Animals, Biology)
-> ungulados -- (Animals, Biology)
-> eutérios, placentários -- (Animals, Biology)
-> mamíferos -- (Animals, Biology)
-> vertebrado -- (Animals, Biology)
-> cordados -- (Animals, Biology)
-> animal, bicho -- (Animals, Biology)
-> criatura, organismo, ser, ser_vivo -- (Biology)
-> coisa, entidade -- (Factotum)

Semantic prototypes vs. Wordnet 2

- tagger/parser-friendly: ideally 1 sem tag, like PoS etc.
- ideally, not more finegrained than what can be disambiguated by linguistic context
  - major classes should allow formal tests or feature contrasting
  - e.g. ±HUM, ±MOVE, type of preposition (“durante”, “em”), ±CONTROL, test-verbs (comer, beber, dizer, produzir)
- careful with “metaphor polysemy explosion”
  - NOT inspired by classical dictionaries
- systematic relations between classes may be left underspecified,
  e.g. <con> --> <unit>, <H> --> <ANIM>, <sport> --> <activity>, <dance> --> <sem-l> <activity>

Lexico-semantic tags in Constraint Grammar

- **secondary**: semantic tags employed to aid disambiguation and syntactic annotation (traditional CG):
  <vcog>, <speak>, <Hprof>, <aloc>, <jnat>
- **primary**: semantic tags as the object of disambiguation
- **existing applications using lexical semantic tags**
  - Named Entity classification (Nomen Nescio, HAREM)
  - semantic prototype tagging for treebanks (Floresta, Arboretum)
  - semantic tag-based applications
    - Machine translation (GramTrans)
    - QA, library IE, sentiment surveys, referent identification (anaphora)

Semantic argument slots

- the semantics of a noun can be seen as a "compromise" between its lexical potential or “form” (e.g. prototypes) and the projection of a syntactic-semantic argument slot by the governing verb (“function”)
- e.g. <civ> (country, town) prototype
  - (a) **location**, **origin**, **destination** slots
    (adverbial argument of movement verbs)
  - (b) **agent** or **patient** slots
    (subject of cognitive or agentive verbs)
- Rather than hypothesize different senses or lexical types for these cases, a **role annotation level** can be introduced as a bridge between syntax and true semantics

Semantic prototypes in the VISL parsers

- ca. 160 types for ~ 35.000 nouns
- SIMPLE- and cross-VISL-compatible (7 languages), thus a possibility for integration across languages
- Ontology with umbrella classes and subcategories, e.g.
  - <H> : <Hprof>, <HH>, <Hnat>, <Htitle>, <Hfam> ...
  - <L> : <Ltop>, <Lh>, <Lwater>, <Labs>, <Lsurf> ....
  - <sem> : <sem-r>, <sem-l>, <sem-c>, <sem-s> ...
- allows composite “ambiguous” tags:
  - <civ> (<HH> + <L>), <media> (<HH> + <sem>)
- metaphors and systematic category inheritance are underspecified: <con> -> <unit>
- prototypes expressed as bundles of atomic semantic features, e.g. <V> (vehicle) = +concrete, -living, -human, +movable, +moving, -location, -time ...

A feature X can be inferred in a given bundle, if there is a feature Y in the same bundle such that – with respect to the whole table - the set of prototype bundles with feature X is a subset of the set of prototype bundles with feature Y.

E = entities (±CONCRETE) V = ±VERBAL
C = ±CONTROL P = ±PERFECTIVE
I = ±MOVING S = ±MEASURING
J = ±MOVABLE D = ±PARTITIVE
A = ±ANIMATE (living) X = ±HUMAN-EXPRESSION (allowing human modifier-ADJ)
H = ±HUMAN ENTITY F = feature (±ADJECTIVAL)
M = ±MASS L = ±LOCATION
N = number (±COUNTABLE) T = ±TEMPORAL

E	C	I	J	A	H	M	N	V	P	S	D	X	F	L	T	
+		'+	+	'+	'+	.	'+			.		+	.	.	.	H, Hprof, Hnat, Hmyth, Hfam, Htit, i, Hbio, Hsick, Hattr, *hum
+		'+	+	'+	'+	.	.			.		+	.	.	.	HH, Hhparty, *party, *media
+			.	.	'+	.	.			.		+	.	'+	.	inst, *inst
+			.	.	'+	.	'+			.		+	.	'+	.	Lciv, *civ
+		'+	+	'+	.	.	'+			.		.	.	.	.	A, Azo, Aorn, Aent, Aich, Amyth, Acell, Adom
+		'+	+	'+	.	.	.			.		.	.	.	.	AA (AAzo ..).
+			.	'+	.	.	'+			.		.	.	.	.	B, Btree, Bbush, Bflo, Bveg
+			.	'+	.	.	.			.		.	.	'+	.	BB
'+			.	.	.	.	'+			.		.	.	'+	.	Ltop, Lwater, Lpath, Lh, Lstar, build, part-build, bar, an, anorg, anost, anzo, anorn,, anbo, sick-c, dir, *top
.			.	.	.	.	'+			.		.	.	'+	.	Labs, Ltip, Lopening, Lsurf
+		.	'+	.	.	.	'+			.		.	.	'+	.	furn, Ltrap, Lcover, con, pict
.			.	.	.	.	.			.		.	.	'+	.	pos-an, pos-soc, sit, anfeat
.			.	.	.	'+	.			.		.	.	'+	.	wea, wea-wind, wea-rain
+		.	'+	.	.	.	'+			.		.	.	'+	.	anmov
'+			.	.	.	.	.			.		.	.	'+	.	Lsurf
+		'+	+	.	.	.	'+			.		.	.	.	.	V, Vground, Vwater, Vair
+		'+	+	.	.	.	.			.		.	.	.	.	VV (ground ....)

prototypes or atomic features?

- *Rioting continued in Paris. The **town** imposed a curfew*
  - anaphoric relations visible as <civ> tag
  - not visible after HUM/PLACE disambiguation
    - Paris +PLACE -HUM, due to “in”
    - town -PLACE +HUM, due to “impose”
- *The Itamarati announced new taxes, but voters may not allow the **government** to go ahead.*
  - semantic context projection (+HUM @SUBJ announce) used to mark metaphorical transfer --> allow reference between the government and its seat (place name)

The disambiguation – metaphor trade-off

- Disambiguation <Azo> vs. <inst>
  - *O **leão** penalizou a especulação*
- Metaphorical re-interpretation of a syntactic slot due to semantic argument projection
  - *O **Itamarati** anunciou novos impostos.*

    <top> *<vH>*

    <+HUM>

    <inst>
- normally head -> dependent (but not exclusively)
  - *um dia triste*
- +HUM overrides -HUM
- concrete --> abstract transfer, not vice versa

Semantic role annotation for Portuguese, Spanish and Danish

- inspired by the Spanish **3LB-LEX** project (Taulé et al. 2005)
- allows, together with the syntactic function annotation (ARG structure) a mapping onto **PropBank** argument frames (Palmer et al. 2005)
- allows the extraction of argument frames from treebanks
- **manual vs. automatic**: due to the quality of the syntactic parser and the existance of the prototype lexicon, a boot-strapping is envisioned, where
  - *syntactic* valency is exploited in conjunction with the prototype lexicon (ontology) to create semantic role annotation,
  - which in turn provides "*semantic* valency frames"
  - which then is used to improve the semantic role annotation

Semantic role granularity

- 52 semantic roles (15 core argument roles and 37 minor and “adverbial” roles)
- Covering the major categories of the tectogrammatical layer of the PDT (Hajicova et al. 2000)
- ARG structure (a la PropBank, Palmer et al 2005) can be added without information loss by combining roles and syntactic function tags
- all clause level constituents are tagged, and where the same categories can be used for group-level annotation, this is annotated, too
- semantic heads: np heads, pp dependents

The semantic role inventory

"Nominal" roles	***definition***	***example***
§AG	agent	X eats Y
§PAT	patient	Y eats X, X broke, X was broken
§REC	receiver	give Y to X
§BEN	benefactive	help X
§EXP	experiencer	*X* fears Y, surprise *X*
§TH	theme	send X, *X* is ill, X is situated there
§RES	result	Y built X
§ROLE	role	Y works as a *guide*
§COM	co-argument, comitative	Y dances with X
§ATR	static attribute	Y is *ill*, a ring o*f gold*
§ATR-RES	resulting attribute	make somebody *nervøs*
§POS	possessor	Y belongs to X, Peter's car
§CONT	content	a bottle of *wine*
§PART	part	Y consists of X, X forms a whole
§ID	identity	the town of *Bergen,* the Swedish company *Volvo*
§VOC	vocative	keep calm, *Peter!*

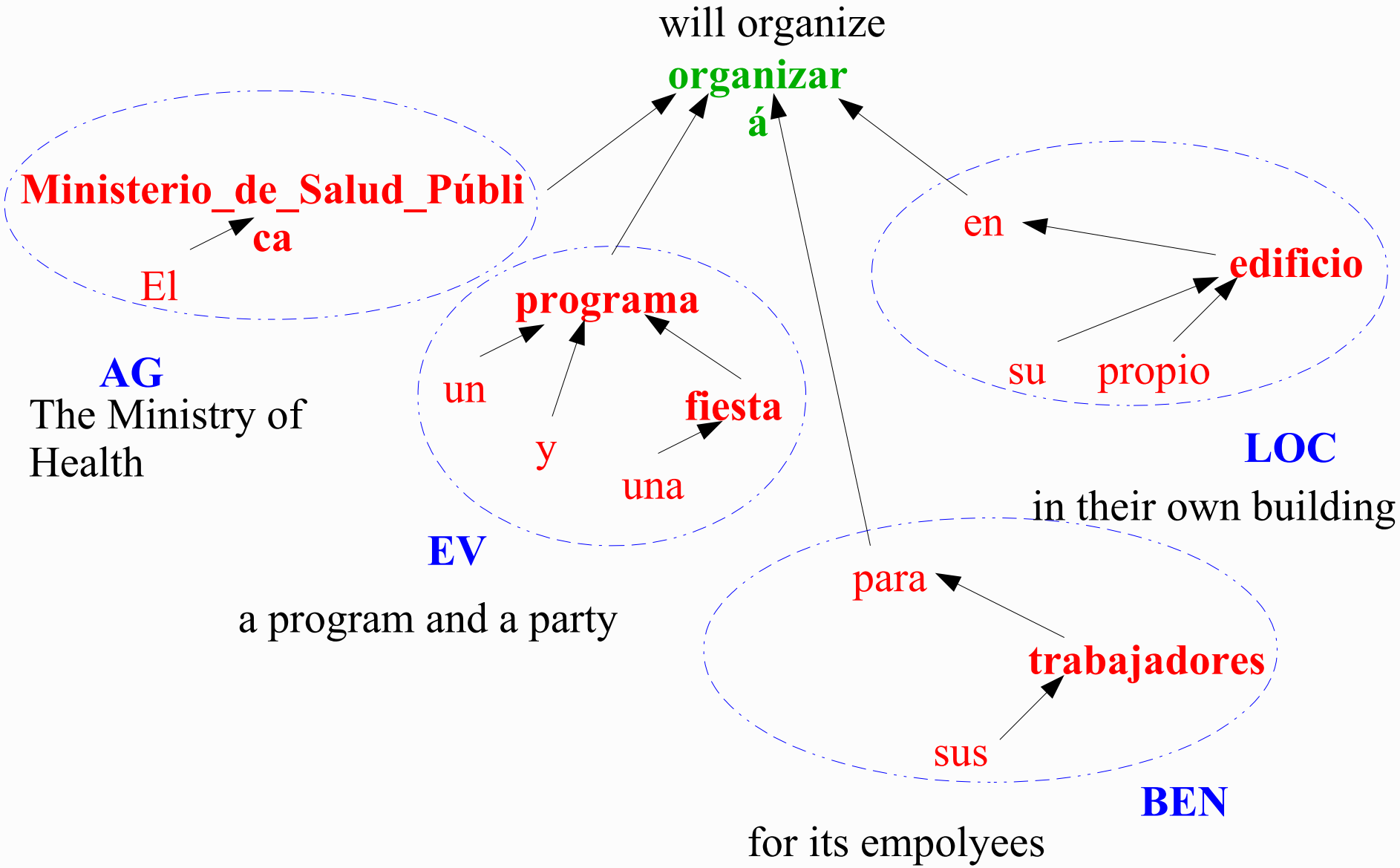
"Adverbial"roles	*definition*	*example*
§LOC	location	live in *X, here, at home*
§ORI	origin, source	flee from *X,* meat from *Argentina*
§DES	destination	send Y *to X,* a flight *to X*
§PATH	path	*down the road, through the hole*
§EXT	extension, amount	march *7 miles,* weigh *70 kg*
§LOC-TMP	temporal location	*last year, tomorrow evening, when we meet*
§ORI-TMP	temporal origin	*since January*
§DES-TMP	temporal destination	*until Thursday*
§EXT-TMP	temporal extension	*for 3 weeks, over a period of 4 years*
§FREQ	frequency	*sometimes, 14 times*
§CAU	cause	*because of X, since he couldn't come himself*
§COMP	comparation	better *than ever*
§CONC	concession	*in spite of X, though we haven't hear anything*
§COND	condition	*in the case of X, unless we are told differently*
§EFF	effect, consequence	*with the result of,* there were som many *that ...*
§FIN	purpose, intention	work *for the ratification of the Treaty*
§INS	instrument	*through* X, cut bread with, come *by car*
§MNR	manner	*this way, as you see fit, how ...*
§COM-ADV	accompanier (ArgM)	*apart from Anne, with s.th. in her hand*

"Syntactic roles"	***definition***	***example***
§META	meta adverbial	*according to X, maybe, apparently*
§FOC	focalizer	*only, also, even*
§ADV	dummy adverbial	if no other adverbial categories apply
§EV	event, act, process	start *X*, ... *X* ends
§PRED	(top) predicatior	main verb in main clause
§DENOM	denomination	lists, headlines
§INC	verb-incorporated	take *place* (not fully implemented)

Exploiting lexical semantic information through syntactic links

- corpus information on verb complementation:
  - CG set definitions
  - e.g. V-SPEAK = “contar” “dizer” “falar” ...
  - MAP (**§SP**) TARGET @SUBJ (**p** V-SPEAK)
- ~ 160 semantic prototypes from the PALAVRAS lexicon
  - e.g. N-LOC = <L> <Ltop> <Lh> <Lwater> <Lparth> <civ> .. combined with destination prepositions PRP-DES = “até” “para” ...
  - MAP (**§DES**) TARGET @P< (0 N-LOC LINK **p** PRP-DES)
- Needs dependency trees as input, created with the syntactic levels of the PALAVRAS parser

Dependency trees



Source format

El *(the)*	[el] <artd> DET @>N **#1->2**
Ministerio=de=Salud=Pública	[M.] <org> PROP M S @SUBJ> **#2->3** $ARG0 **§AG**
organizará *(organized)*	[organizar] <aux> V FUT 3S IND @FS-STA **#3->0** **§PRED**
un *(a)*	[un] <arti> DET M S @>N **#4->5**
programa *(program)*	[programa] <cjt-head> <act> N M S @<ACC **#5->3** $ARG1 **§EV**
y *(and)*	[y] <co-acc> KC @CO **#6->5**
una *(a)*	[un] <arti> DET F S @>N **#7->8**
fiesta *(party)*	[fiesta] <cjt> <act> N M S @<ACC **#8->5** $ARG1 **§EV**
para *(for)*	[para] PRP @<ADVL **#9->3**
sus *(their)*	[su] <poss> <si> DET M P @>N @>N **#10->11**
trabajadores *(workers)*	[trabajador] <Hprof> N M P @P< **#11->9** **§BEN**
en *(in)*	[en] PRP @<ADVL **#12->3**
su *(their)*	[su] <poss> <si> DET M S @>N @>N **#13->15**
propio *(own)*	[propio] <ident> DET M S @>N @>N **#14->15**
edifício *(building)*	[edifício] <build> N M S @P< **#15->12** **§LOC**

(authentic newspaper text)

Inferring **semantic roles** from **verb classes** **and** **syntactic function** (@) and dependency (p, c and s)

implicit inference of semantics:

syntactic function (e.g. @SUBJ) and valency potential (e.g. ditransitive <vdt>) are *not* semantic by themselves, but help restrict the range of possible argument roles (e.g. §BEN for @DAT)

- *subjects of ergatives*
  MAP (§PAT) TARGET @SUBJ (p <ve> LINK NOT c @ACC) ;
- *the give sb-DAT s.th.-ACC frame*
  MAP (§TH) TARGET @ACC (s @DAT) ;

Inferring **semantic roles** from **semantic prototype sets** using syntactic function (@) and dependency (p, c and s)

explicit use of lexical semantics: semantic prototypes: <Hprof> (human professional),
<Hideo> (ideology-follower),
<Hnat> (nationality) ...
restrict the role range by themselves, but are ultimately still dependent on verb argument frames

(a) "*Genitivus objectivus/subjectivus*"

- MAP (§PAT) TARGET @P< (p PRP-AF + @N< LINK p N-VERBAL) ;

  # the destruction of the town
- MAP (§AG) TARGET GEN @>N (p N-ACT) ;
  # The government's release of new data
- MAP (§PAT) TARGET GEN @>N (p N-HAPPEN) ;
  # The collapse of the economy

- **Agent**: "he was chased by three police cars"
  MAP (§AG) TARGET @P< (p ("by" @ADVL) LINK p PAS) (0 N-HUM OR N-VEHICLE) ;
- **Possessor**: "the painter's brush"
  MAP (§POS) TARGET @P< (0 N-HUM + GEN LINK 0 @>N) (p N-OBJECT) ;
- **Instrumental**: “destroy the piano with a hammer”
  MAP (§INS) TARGET @P< (0 N-TOOL) (p ("with") + @ADVL) ;
- **Content**: “a bottle of wine"
  MAP (§CONT) TARGET @P< (0 N-MASS OR (N P)) (p ("of") LINK p **<con>**) ;
- **Attribute**: “a statue of gold”
  MAP (§ATR) TARGET @P< + N-MAT (p ("of") + @N<) ;
- **Location**: “live in a big house”
  MAP (§LOC) TARGET @P< + N-LOC (p PRP-LOC LINK 0 @ADVL OR @N<);
- **Origin**: “send greetings from Athens”, “drive all the way from the border”
  MAP (§ORI) TARGET @P< (0 N-LOC) (p PRP-FROM LINK 0 @<ADVL OR @<SA OR @<OA LINK p V-MOVE/TR) ;
- **Temporal extension**: “The session lasted 4 hours”
  MAP (§EXT-TMP) TARGET @SA (0 N-DUR) ;

Semantic role tagging performance on CG-revised Floresta + live dependency + live prototype tagging
R=86.8%, P=90.5%, F=88.6%

role label		*recall*	*precision*	*F-Score*
§FOC	t	97.4 %	97.4 %	97.4
§REFL	t	100 %	94.7 %	97.3
§DENOM	t	100 %	93.8 %	96.8
§PRED	t	97.4 %	96.1 %	96.7
§ATR	C, np	91.7 %	97.7 %	94.5
§ID	np	100 %	93.3 %	90.6
§AG	C	92.7 %	87.4 %	90.0
§PAT	C	91.5 %	86.6 %	89.0
§LOC	C	92.0 %	76.7 %	88.9
§ORI	C	100 %	80.0 %	88.9
all categories		**86.6 %**	**90.5 %**	**88.6**
§TH	C	81.6 %	86.6 %	84.0
§FIN	a	79.2 %	86.4 %	81.7
§LOC-TMP	a	87.1 %	72.8 %	79.3
§CAU	a	86.7 %	72.2 %	78.8
§RES	C	74.1 %	83.3 %	78.4
§BEN	C	80.0 %	72.7 %	76.2
§DES	C	84.6 %	68.8 %	75.9
§ADV	a	100 %	57.9 %	72.2

Corpus results from a recent Spanish sister project

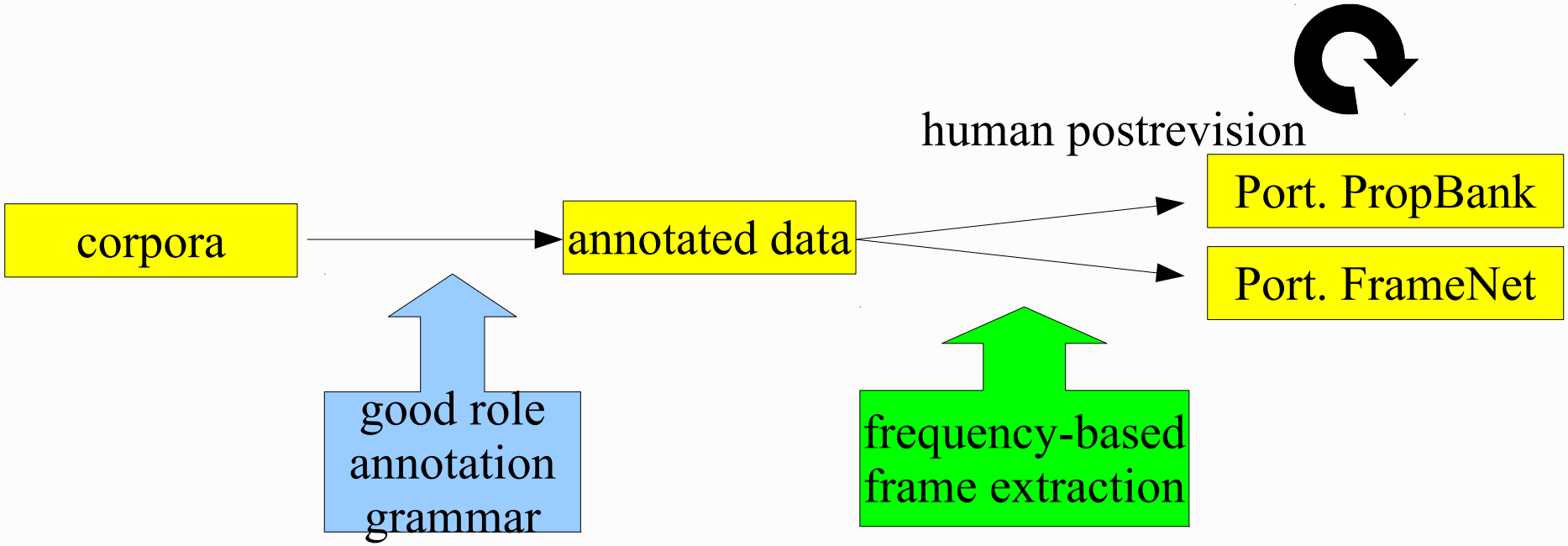
- compilation and annotation of a Spanish internet corpus (11.2 million words)
- to infer tendencies about the relationship between semantic roles and other grammatical categories:

Role	Syntactic function[1]	Part of speech[2]	Semantic prototype[3]
§TH	ACC (61%)	N (57%)	sem-c (10%)
§AG	SUBJ> (91%)	N (45%)	Hprof (7%)
§ATR	SC (75%)	N, ADJ, PCP	act (7%)
§BEN	ACC (55%)	INDP (35%)	HH (13%)
§LOC-TMP	ADVL (64%)	ADV (34%)	per (31%)
§EV	ACC (54%)	N (85%)	act (33%)
§LOC	ADVL (57%)	PRP-N (55%)	L (10%)
§REC	DAT (73%)	PERS (41%)	H (9%)
§TP	FS-ACC (34%)	VFIN (33%)	sem-c (14%)
§PAT	SUBJ> (73%)	N (55%)	sem-c (7%)

- smallest syntactic “spread”: §AG, §COG, §SP (subject and agent of passive)
- easy: §SP and §COG, inferable from the verb alone
- difficult: §TH, covers a wide range of verb types and semantic features
- @SUBJ and @ACC match >= 20 roles, but unevenly
- human roles tend to appear left, others right

Role	Frequency	Subject/object ratio	Left/Right ratio
§TH	14.6 %	25.4 %	31.0 %
§AG	6.6 %	97.2 %	78.4 %
§ATR	6.0 %	-	21.7 %
§BEN	5.0 %	3.2 %	59.2 %
§LOC-TMP	4.0 %	23.7 %	42.6 %
§EV	3.7 %	43.4 %	30.0 %
§LOC	3.0 %	0.0 %	23.0 %
§REC	1.6 %	87.8 %	44.7 %
§TP	1.5 %	4.0 %	7.5 %
§PAT	0.4 %	80.0 %	68.5 %

- Problems
  - interdependence between syntactic and semantic annotation
  - multi-dimensionality of prototypes (e.g. <coll>, <part>, <group>)
  - a certain gradual nature of role definitions
  - the verb frame bottleneck
- Plans:
  - annotate what is possible, one argument at a time, use function generalisation and noun types where verb frames are not available
  - Boot-strap a frame lexicon from automatically role-annotated text



VISL

http://beta.visl.sdu.dk
http://corp.hum.sdu.dk
http://www.gramtrans.com/deepdict/

eckhard.bick@mail.dk

****

DeepDict-generated stub sentences as prototypical, semantics-defining usage examples

- alien allegedly **abducts** child
- PROP/act effectively **abolishes** slavery
- PROP/commission gratefully **accepts** amendment | on behalf | at university | to extent | under circumstance | without reservation | within framework
- PROP/bowler successfully **accomplishes** feat
- problem: polysemy interference when using only binary relations
  - sediment consciously **accumulates** wealth | in cell | over time | as consequence
  - PROP/album **sells** goods | to devil | at price | into slavery | for scrap | under name | as slave | in exchange | on market | without license
- problem: surface polishing: article insertion, singular/plural decision, PROP-typing