Annotating Corpora for Linguistics
from text to knowledge

Eckhard Bick
University of Southern Denmark
Research advantages of using a corpus rather than introspection

- **empirical, reproducible**: 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

- 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
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 bootstrapping, 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"?)

- **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

Descriptive
Motivation: Explanatory
Test: teaching

Methodological
Robust
Machine translation
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:

- **Topological** 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?

VISL grammar games:

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

OrdRet

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

NER

- Annotated corpora: CorpusEye

QA
## CG languages (VISL/GS)

<table>
<thead>
<tr>
<th></th>
<th>Parser</th>
<th>Lexicon</th>
<th>Grammar</th>
<th>Appl.</th>
<th>Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>🇩🇰</td>
<td>DanGram</td>
<td>100,000 lexemes, 40,000 names</td>
<td>8,400 rules</td>
<td>MT, grammar checker, NER, teaching, QA</td>
<td>ca. 150 mill. words (mixed, news)</td>
</tr>
<tr>
<td>🇵🇹</td>
<td>PALAVRAS</td>
<td>70,000 lexemes, 15,000 names</td>
<td>7,500 rules</td>
<td>teaching, NER, QA, MT</td>
<td>ca. 380 mill. words (news, wiki, europarl ...)</td>
</tr>
<tr>
<td>🇪🇸</td>
<td>HISPAL</td>
<td>73,000 lexemes</td>
<td>4,900 rules</td>
<td>teaching</td>
<td>ca. 86 mill. words (Wiki, Europarl, Internet)</td>
</tr>
<tr>
<td>🇬🇧</td>
<td>EngGram</td>
<td>81,000 val/sem</td>
<td>4,500 rules</td>
<td>teaching, MT</td>
<td>ca. 210 mill. words (mixed)</td>
</tr>
<tr>
<td>🇸🇪</td>
<td>SweGram</td>
<td>65,000 val/sem</td>
<td>8,400 rules</td>
<td>teaching, MT</td>
<td>ca. 60 mill. words (news, Europarl, wiki)</td>
</tr>
<tr>
<td>🇳🇴</td>
<td>NorGram</td>
<td>OBT / via DanGram</td>
<td>OBT / via DanGram</td>
<td>teaching, MT</td>
<td>ca. 30 mill. words (Wikipedia)</td>
</tr>
<tr>
<td>🇫🇷</td>
<td>FrAG</td>
<td>57,000 lexemes</td>
<td>1,400 rules</td>
<td>teaching</td>
<td>67 mill (Wiki, Europarl)</td>
</tr>
<tr>
<td>🇩🇪</td>
<td>GerGram</td>
<td>25,000 val/sem</td>
<td>ca. 2000 rules</td>
<td>teaching, MT</td>
<td>ca. 44 mill. words (Wiki, Europarl, mixed)</td>
</tr>
<tr>
<td>🇪🇸</td>
<td>EspGram</td>
<td>30,000 lexemes</td>
<td>2,600 rules</td>
<td>grammar checker, MT</td>
<td>ca. 40 mill. words (mixed, literature, internet, news)</td>
</tr>
<tr>
<td>🇮🇹</td>
<td>ItaGram</td>
<td>30,600 lexemes</td>
<td>1,600 rules</td>
<td>teaching</td>
<td>46 mill. (Wiki, Europarl)</td>
</tr>
</tbody>
</table>
VISL languages (others)

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

Primary tags:
- Pos
- morphology
- @function
- %roles
- #n->m relations

Lexical secondary tags:
- valency: <vt>, <vi>, <+on>
- semantic class: <atemp>
- semantic prototype: <tool>

Functional secondary tags:
- verb chain: <aux>, <mv>
- attachment: <np-close>
- coordinator function: <co-fin>
- clause boundaries: <clb> <break>

clause boundaries:
- <clb> <break>
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 it 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)) ;
Cohorts

“<sails>”
“sail” V PR 3S
“sail” N P NOM

Disambiguation

Syntax

Mapping

Substitution

polysemy

sem. roles

Disambiguation

Mapping

Dep.

PSG

extenal modules

CG flowchart

TEXT

Analyzer

Lexica

Morphology

External e.g. DTT
tagger
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: 3,046
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

Text Painter

Languages: Danish, English, Esperanto, French, German, Portuguese, Spanish

Subjects: direct/accusative objects, indirect/dative objects, adverbials (free or bound)

Nouns: proper nouns, adjectives, adverbs

Enter text to parse:

træk, hvor man kan bruge sin egen tekst. Hvis bruger vil teste sig selv, kan han bruge Text Painter interaktivt.

Parser: Standard Parser

Visualization: Selected category highlight

Categories: @SUBJ ... OR ... NONE

Text=Painter er et redskab til visualisering af grammatiske træk, hvor man kan bruge sin egen tekst. Hvis brugeren vil teste sig selv, kan han bruge Text=Painter interaktivt
WebPainter

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

mouse-over translation:

```
All of the above bicycle races involve diamond frame bicycles of two triangles. An alternative is the recumbent, a bicycle on which the rider sits back with the legs horizontal. This puts the body in a position where there is less wind drag. Proponents claim it provides more comfortable riding, with no weight on the wrists. The recumbent is a more aerodynamic design of bicycle, and world speed records were set with them.
```

optional grammar (here: SUBJ and prep)

```
Bicycle races are popular all over the world, especially in Europe. The most devoted countries are Italy, Spain, Belgium, Germany, France, the Netherlands and Switzerland, although the United States has international standing, as does Australia. The USA boasts three-time Tour de France winner and first American winner, Greg LeMond as well as seven-time winner Lance Armstrong. Australia has seen success through Michael Rogers (World Road Time Trial Champion, 2003, 2004, and...
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 news text, ~ 60M words annotated
- **DFK**, mainly transcribed 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**, transcribed 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

- (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)

e.g. “Washington”

+LOC  

<hum> (person)

-HUM  

<hum> (person)

+civ> (town, country)
Semantic prototypes vs. Wordnet

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

. cavalo -- (Animals, Biology)
  -> equídeos -- (Animals, Biology)
  -> perissodáctilos -- (Animals, Biology)
  -> ungulados -- (Animals, Biology)
  -> eutéritos, 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>`, `<alloc>`, `<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 $Y$. 

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

<table>
<thead>
<tr>
<th>&quot;Nominal&quot; roles</th>
<th>definition</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>§AG agent</td>
<td>X eats Y</td>
<td></td>
</tr>
<tr>
<td>§PAT patient</td>
<td>Y eats X, X broke, X was broken</td>
<td></td>
</tr>
<tr>
<td>§REC receiver</td>
<td>give Y to X</td>
<td></td>
</tr>
<tr>
<td>§BEN benefactive</td>
<td>help X</td>
<td></td>
</tr>
<tr>
<td>§EXP experiencer</td>
<td>X fears Y, surprise X</td>
<td></td>
</tr>
<tr>
<td>§TH theme</td>
<td>send X, X is ill, X is situated there</td>
<td></td>
</tr>
<tr>
<td>§RES result</td>
<td>Y built X</td>
<td></td>
</tr>
<tr>
<td>§ROLE role</td>
<td>Y works as a guide</td>
<td></td>
</tr>
<tr>
<td>§COM co-argument, comitative</td>
<td>Y dances with X</td>
<td></td>
</tr>
<tr>
<td>§ATR static attribute</td>
<td>Y is ill, a ring of gold</td>
<td></td>
</tr>
<tr>
<td>§ATR-RES resulting attribute</td>
<td>make somebody nervøs</td>
<td></td>
</tr>
<tr>
<td>§POS possessor</td>
<td>Y belongs to X, Peter's car</td>
<td></td>
</tr>
<tr>
<td>§CONT content</td>
<td>a bottle of wine</td>
<td></td>
</tr>
<tr>
<td>§PART part</td>
<td>Y consists of X, X forms a whole</td>
<td></td>
</tr>
<tr>
<td>§ID identity</td>
<td>the town of Bergen, the Swedish company Volvo</td>
<td></td>
</tr>
<tr>
<td>§VOC vocative</td>
<td>keep calm, Peter!</td>
<td></td>
</tr>
<tr>
<td>&quot;Adverbial&quot; roles</td>
<td>definition</td>
<td>example</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>§LOC</td>
<td>location</td>
<td>live in X, here, at home</td>
</tr>
<tr>
<td>§ORI</td>
<td>origin, source</td>
<td>flee from X, meat from Argentina</td>
</tr>
<tr>
<td>§DES</td>
<td>destination</td>
<td>send Y to X, a flight to X</td>
</tr>
<tr>
<td>§PATH</td>
<td>path</td>
<td>down the road, through the hole</td>
</tr>
<tr>
<td>§EXT</td>
<td>extension, amount</td>
<td>march 7 miles, weigh 70 kg</td>
</tr>
<tr>
<td>§LOC-TMP</td>
<td>temporal location</td>
<td>last year, tomorrow evening, when we meet</td>
</tr>
<tr>
<td>§ORI-TMP</td>
<td>temporal origin</td>
<td>since January</td>
</tr>
<tr>
<td>§DES-TMP</td>
<td>temporal destination</td>
<td>until Thursday</td>
</tr>
<tr>
<td>§EXT-TMP</td>
<td>temporal extension</td>
<td>for 3 weeks, over a period of 4 years</td>
</tr>
<tr>
<td>§FREQ</td>
<td>frequency</td>
<td>sometimes, 14 times</td>
</tr>
<tr>
<td>§CAU</td>
<td>cause</td>
<td>because of X, since he couldn't come himself</td>
</tr>
<tr>
<td>§COMP</td>
<td>comparation</td>
<td>better than ever</td>
</tr>
<tr>
<td>§CONC</td>
<td>concession</td>
<td>in spite of X, though we haven't hear anything</td>
</tr>
<tr>
<td>§COND</td>
<td>condition</td>
<td>in the case of X, unless we are told differently</td>
</tr>
<tr>
<td>§EFF</td>
<td>effect, consequence</td>
<td>with the result of, there were som many that ...</td>
</tr>
<tr>
<td>§FIN</td>
<td>purpose, intention</td>
<td>work for the ratification of the Treaty</td>
</tr>
<tr>
<td>§INS</td>
<td>instrument</td>
<td>through X, cut bread with, come by car</td>
</tr>
<tr>
<td>§MNR</td>
<td>manner</td>
<td>this way, as you see fit, how ...</td>
</tr>
<tr>
<td>§COM-ADV</td>
<td>accompanier (ArgM)</td>
<td>apart from Anne, with s.th. in her hand</td>
</tr>
<tr>
<td>&quot;Syntactic roles&quot;</td>
<td>definition</td>
<td>example</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>§META</td>
<td>meta adverbial</td>
<td>according to $X$, maybe, apparently</td>
</tr>
<tr>
<td>§FOC</td>
<td>focalizer</td>
<td>only, also, even</td>
</tr>
<tr>
<td>§ADV</td>
<td>dummy adverbial</td>
<td>if no other adverbial categories apply</td>
</tr>
<tr>
<td>§EV</td>
<td>event, act, process</td>
<td>start $X$, ... $X$ ends</td>
</tr>
<tr>
<td>§PRED</td>
<td>(top) predicator</td>
<td>main verb in main clause</td>
</tr>
<tr>
<td>§DENOM</td>
<td>denomination</td>
<td>lists, headlines</td>
</tr>
<tr>
<td>§INC</td>
<td>verb-incorporated</td>
<td>take place (not fully implemented)</td>
</tr>
</tbody>
</table>
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
The Ministry of Health will organize a program and a party for its employees in their own building.
El (the) Ministerio=de=Salud= Pública organizará (organized) un (a) programa (program) y (and) una (a) fiesta (party) para (for) sus (their) trabajadores (workers) en (in) su (their) propio (own) edificio (building)
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%

<table>
<thead>
<tr>
<th>role label</th>
<th>recall</th>
<th>precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>§FOC</td>
<td>t</td>
<td>97.4%</td>
<td>97.4%</td>
</tr>
<tr>
<td>§REFL</td>
<td>t</td>
<td>100%</td>
<td>94.7%</td>
</tr>
<tr>
<td>§DENOM</td>
<td>t</td>
<td>100%</td>
<td>93.8%</td>
</tr>
<tr>
<td>§PRED</td>
<td>t</td>
<td>97.4%</td>
<td>96.1%</td>
</tr>
<tr>
<td>§ATR</td>
<td>C, np</td>
<td>91.7%</td>
<td>97.7%</td>
</tr>
<tr>
<td>§ID</td>
<td>np</td>
<td>100%</td>
<td>93.3%</td>
</tr>
<tr>
<td>§AG</td>
<td>C</td>
<td>92.7%</td>
<td>87.4%</td>
</tr>
<tr>
<td>§PAT</td>
<td>C</td>
<td>91.5%</td>
<td>86.6%</td>
</tr>
<tr>
<td>§LOC</td>
<td>C</td>
<td>92.0%</td>
<td>76.7%</td>
</tr>
<tr>
<td>§ORI</td>
<td>C</td>
<td>100%</td>
<td>80.0%</td>
</tr>
<tr>
<td>all categories</td>
<td></td>
<td>86.6 %</td>
<td>90.5 %</td>
</tr>
<tr>
<td>§TH</td>
<td>C</td>
<td>81.6%</td>
<td>86.6%</td>
</tr>
<tr>
<td>§FIN</td>
<td>a</td>
<td>79.2%</td>
<td>86.4%</td>
</tr>
<tr>
<td>§LOC-TMP</td>
<td>a</td>
<td>87.1%</td>
<td>72.8%</td>
</tr>
<tr>
<td>§CAU</td>
<td>a</td>
<td>86.7%</td>
<td>72.2%</td>
</tr>
<tr>
<td>§RES</td>
<td>C</td>
<td>74.1%</td>
<td>83.3%</td>
</tr>
<tr>
<td>§BEN</td>
<td>C</td>
<td>80.0%</td>
<td>72.7%</td>
</tr>
<tr>
<td>§DES</td>
<td>C</td>
<td>84.6%</td>
<td>68.8%</td>
</tr>
<tr>
<td>§ADV</td>
<td>a</td>
<td>100%</td>
<td>57.9%</td>
</tr>
</tbody>
</table>
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:

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

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

human postrevision

Port. PropBank
Port. FrameNet

good role annotation grammar
frequency-based frame extraction

annotated data

corpora
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