Eckhard Bick

VISL
An Integrated Multi-lingual Approach to ICALL

Talk outline

- Background: VISL project activities
- A unified approach to grammar teaching
- Internet based teaching tools
- Grammar Games
- TextPainter: Visualising grammatical text properties
- Research corpora: A ressource for teaching
- Slot filler exercises: Towards evaluation

Teaching projects

- **CTU** 1996-99: Internet based grammar teaching software (research and development)
- **ELU1** 1998-2000: VISL tools for Danish universities and teacher seminaries
- **VISL-HHX** 2001-03: VISL tools for Danish business schools
- **VISL-GYM** 2001-02: VISL tools for Danish gymnasiums
- **PaNoLa, GREI** 2002-2004: Major Nordic languages
- **VISL-SEM** 2004-05: VISL didactics for teacher training colleges
- **URKAS** 2004-05: Language awareness (1.g)

Unity in diversity:
A unified approach for 25 languages

Advantages of the multi-lingual unified approach

- Pooling of teaching time ressources across languages, and even across grades
- Terminological facilitation: stable terms & abbreviations
- Language awareness: direct structural and lexical comparisons across languages
- Shared technology: Games, Corpus searches, ...
- Shared meta-information: texts, exercises, didactics: “accidental” funding or teacher contributions can easily be shared by others

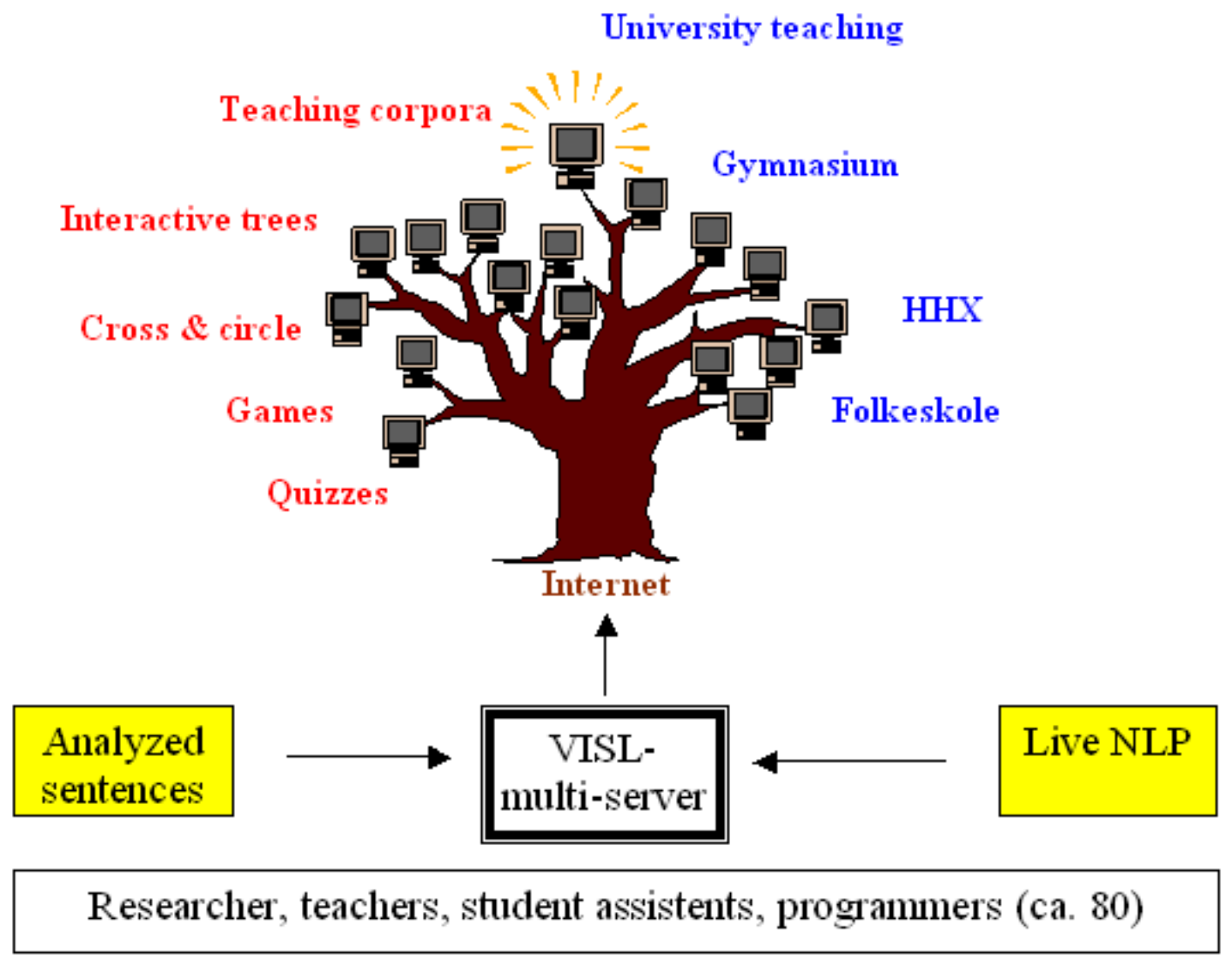
NLP support

- Parsers as a pre-stage for revised analyses (treebanks): more material for less money
- Language awareness: compilation, annotation and search interfaces for (text) corpora
- Explorative use of structural analysis, text type visualisation, category statistics
- Text-independence: any textbook, any quote, any made-up sentence can be incorporated (either revised or live)
- Teacher's angle: Finding examples
- Discussion errors: Grammar checking, MT

VISL research languages & treebank tools

	revised syntactic trees (tokens)	morphological analysis	syntactic analysis	semantics
	200.000* 4 subcorpora	lexicon and rule based analyzer + CG	CG + DEP	semantic prototypes Po-Da MT, NER
	40.400 13 subcorpora	integrated TWOL/CG (lingsoft) + add-on	CG + PSG or DEP	WordNet based tagging
	425.000* 9 subcorpora	lexicon and rule based analyzer + CG	CG + PSG or DEP or topol.	semantic prototypes Da-En/Eo MT, NER
	8.400 3 subcorpora	lexicon and rule based analyzer + CG	CG + tree-generator	-
	16.000 3 subcorpora	integrated TWOL/CG (lingsoft) + add-on	CG + PSG	semantic prototypes (experimental)
	30.000 4 subcorpora	Decision Tree Tagger (H.Schmid & A.Stein)	CG + PSG or DEP	-
	1.000 2 subcorpora	Decision Tree Tagger (H.Schmid & A.Stein)	CG	-
	-	morpheme based analyzer + CG	CG (experimental)	Da-Esp MT

The VISL teaching network



Warschauer:	Behaviouristic	Communicative	Integrational
Cognitive style favoured	behaviourism field-independent	assimilation field-dependent	cognitivism, conceptual differentiation
Learning	explicit & route learning drill & practice assessment	implicit (inter)active discussion-based	explorative language awareness
Human dimension	individual	social, direct	global, remote
Tools, hardware	single school PC/screen	shared/home PC home PC, CD-ROM	networked PC DVD
Tools, software	hot potatoes: slot filler, matching & completion exercises multiple choice	simulated environment spellcheckers, simple concordances, games (competition/ highscores)	full NLP, some MT grammar checkers annotated corpora games
Language	text book language	productive, simulated communicative	live comm. (e.g. chat, e-mail), multi-genre
Media	text computer as a versatile variety paper	beginning multimedia (speech production, graphics, cd-rom)	full multimedia (video, speech recognition) internet
Information	static	interactive/cooperative information handling	generalized dynamic

Placing VISL

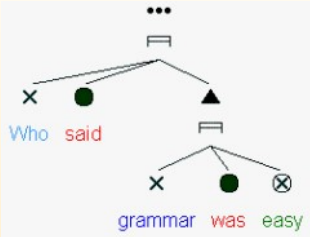
	Behaviouristic	**Communicative**	**Integrational**
Learning	explicit & route learning drill & practice [assessment]	prototype: AnimalQuiz	explorative language awareness (URKAS)
Tools, hardware	user-side java & javascript	- [no videoconferencing]	internet interface remote database access
Tools, software	hot potatoes KillerFiller	games (competition, highscores): WordFall, Labyrinth, SpaceRescue AnimalQuiz	live tree analysis TextPainter Grammar-checker some MT search interfaces statistics
Language	text book examples pedagogical treebanks	Grammy Story Line [no live orspoken communication]	real-life corpora, including chat, e-mail 26 languages with unified descriptive system
Media	on-line teaching texts	graphics some sound some comments	internet [no speech recognition] [no video clips]

A unified descriptive system for 25 languages: Function & form

- The VISL cafeteria of categories
  - Functions: S, P, Od, Oi, Op, Cs, Co, A ...
  - Forms:
    - Complex: cl (clause), g (group), par (paratagma)
    - Simple: n (noun), v (verb), adj, adv, prp, ...
- Pedagogical conventions
  - Constituent trees for teaching, dependency for research
  - No non-branching non-terminal nodes, conventions about ellipsis, zero-constituents, discontinuity ...

Function categories

●		Predicator (P), Verbal (V)
	○	Auxiliary (Vaux), also as <> (D)
	✪	Main verb (V*, Vm), also as ★ (H, K)
	⦿	Verb chain particle (Vpart), also as <> (D), simplified as ⮟(A)
	➲	Infinitive marker (Vi, INFM), also as <> (D)
✖		Subject (S)
	(✖)	Formal or provisional subject (Sf), possibly with the subclass of situative subject (Ss)
▲		Direct (accusative) object (Od)
	(▲)	Formal or provisional object (Of)
■		Indirect (dative) object (Oi)
◆		Prepositional object (Op), at the lite level filtered into ⮟(A)
⊗		Subject complement (Cs), Subject predicative (Ps)
	[⊗]	Free subject predicative (fPs, fCs), simplified as ⮟(A)
⊕		Object complement (Co), Object predicative (Po)
	[⊕]	Free object predicative (fPo, fCo), simplified as ⮟(A)
⮟		Adverbial (A), with possible subdivision of free (⮟ fA) or bound (⮝ bA, bAs, bAo)
★		Head (H), Kernel (K)
<>		Dependents (D)
⬊		Subordinator (SUB)
↔		Co-ordinator (CO)
#		Conjunct (CJT)



Choose tool	e.g. **inspection**, **build tree** or **label tree**
Choose complexity	e.g. **minor** (dynamic sentence dependent reduction in category complexity) or **major**
Choose notation	e.g. **symbols** or **abbrebiations** and/or **colors**
Choose teaching environment	e.g. **latinate Danish gymnasium**
Choose meta-language	e.g. **English**
Choose visualisation	e.g. **graphical trees** or **field analysis**
Choose level	e.g. **VISL-lite** (for schools)
Choose subcorpus	e.g. **VISL-HHX** (business gymnasium)
Choose target language	e.g. **German** or **Swedish**

Teaching corpora of analyzed sentences

Complexity progression

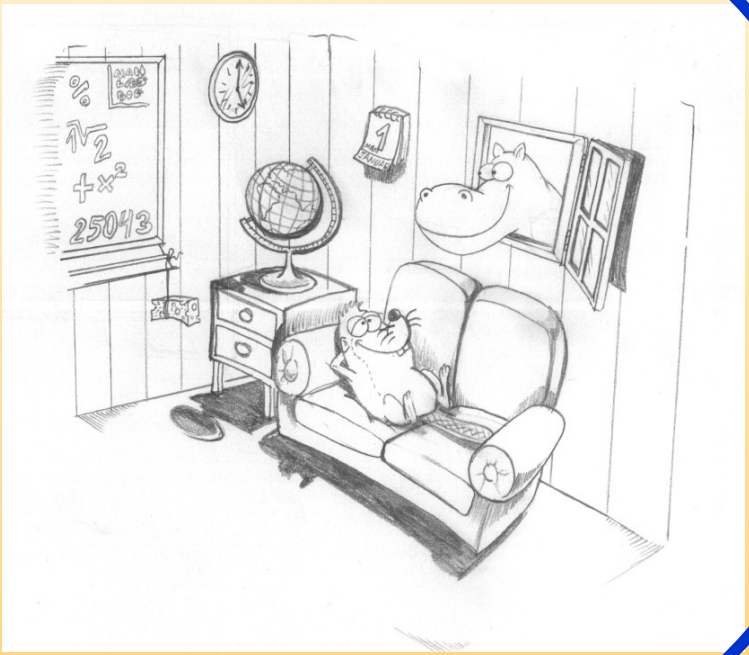
Topic	Formalism	Method
word classes 1 (PoS) optional: morphology	**PoS color-coding** optional: inflexion endings	1. black board-introduction, underlining, *match form/function* 2. *Paintbox* game (initially reduced PoS set) 3. *ShootingGallery, WordFall* 4. *Labyrinth* (later, in syntactic phase) optional: morphology game *(Balloons)*
SVO functions (2) later: adverbials / predicatives	word-based **cross & circle** optional: case marking	1. black board-introduction, cross & circle word level 2. *Postoffice* game (initially reduced category set)
phrases/groups (5) heads & dependents	phrase-based **cross & circle**, simple trees	1. Cross & circle constituent level (underlining) 2. *Java SyntaxTrees* (inspection): lite & minor
coordination (6) verb groups (7)	**syntactic tree structures**	1. "flat"/word-based: *Postoffice* game 2. deep/group-based: *Java SyntaxTrees* (inspection)
subclauses (8) infinitives (9) punctuation rules	complex trees	1. *Java SyntaxTrees* (inspection): lite & major 2. *SynTris* game 3. *SpaceRescue* game 4. *Java SyntaxTrees* (interactive tree-building)
live sentences	unorthodox trees	1. *Java SyntaxTrees:* default & major

Grammy i Klostermølleskoven

9 levels/topics

Story-line about grammar

Comments for teachers

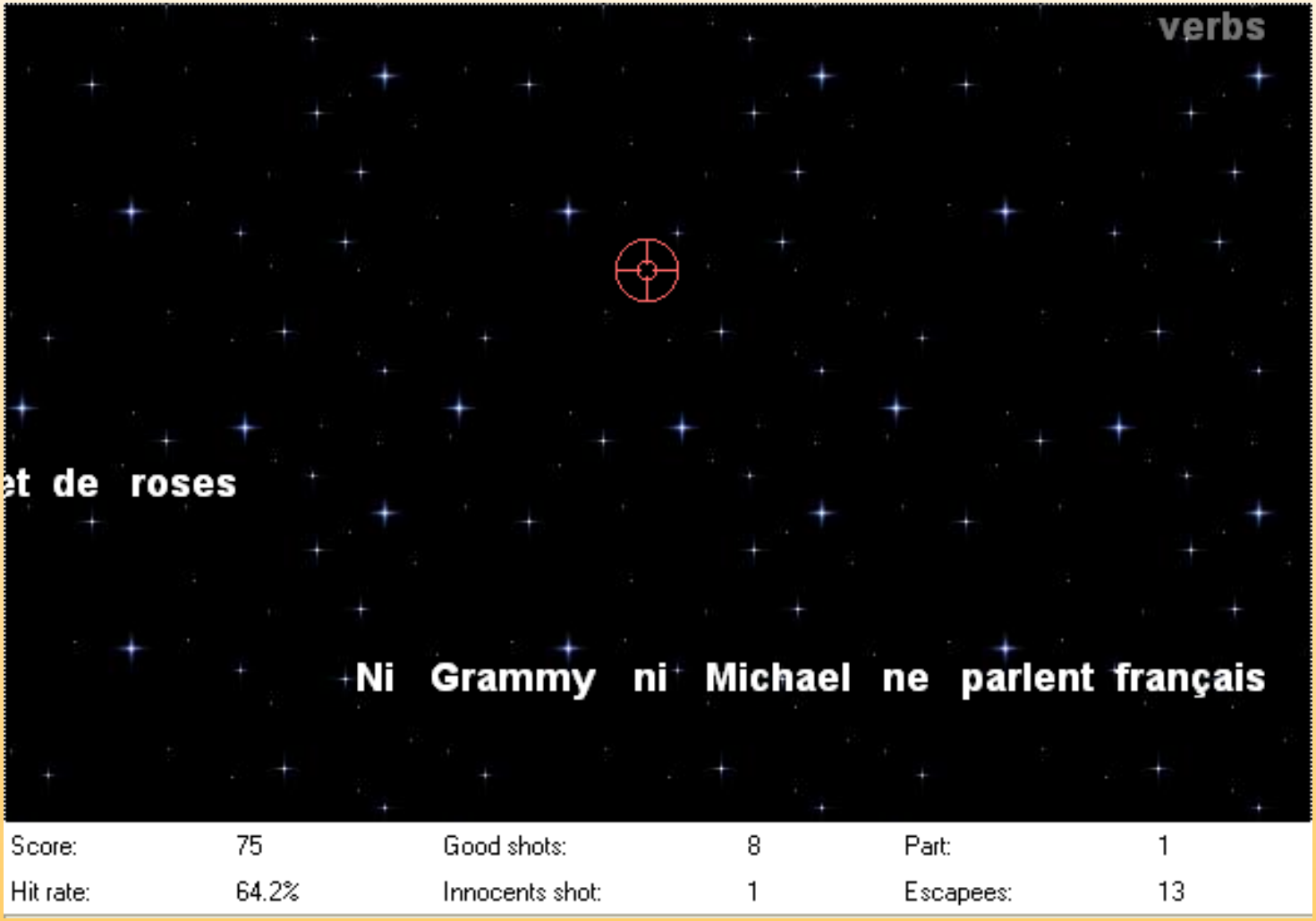


Interactive exercises Book = IT

Explanations for students

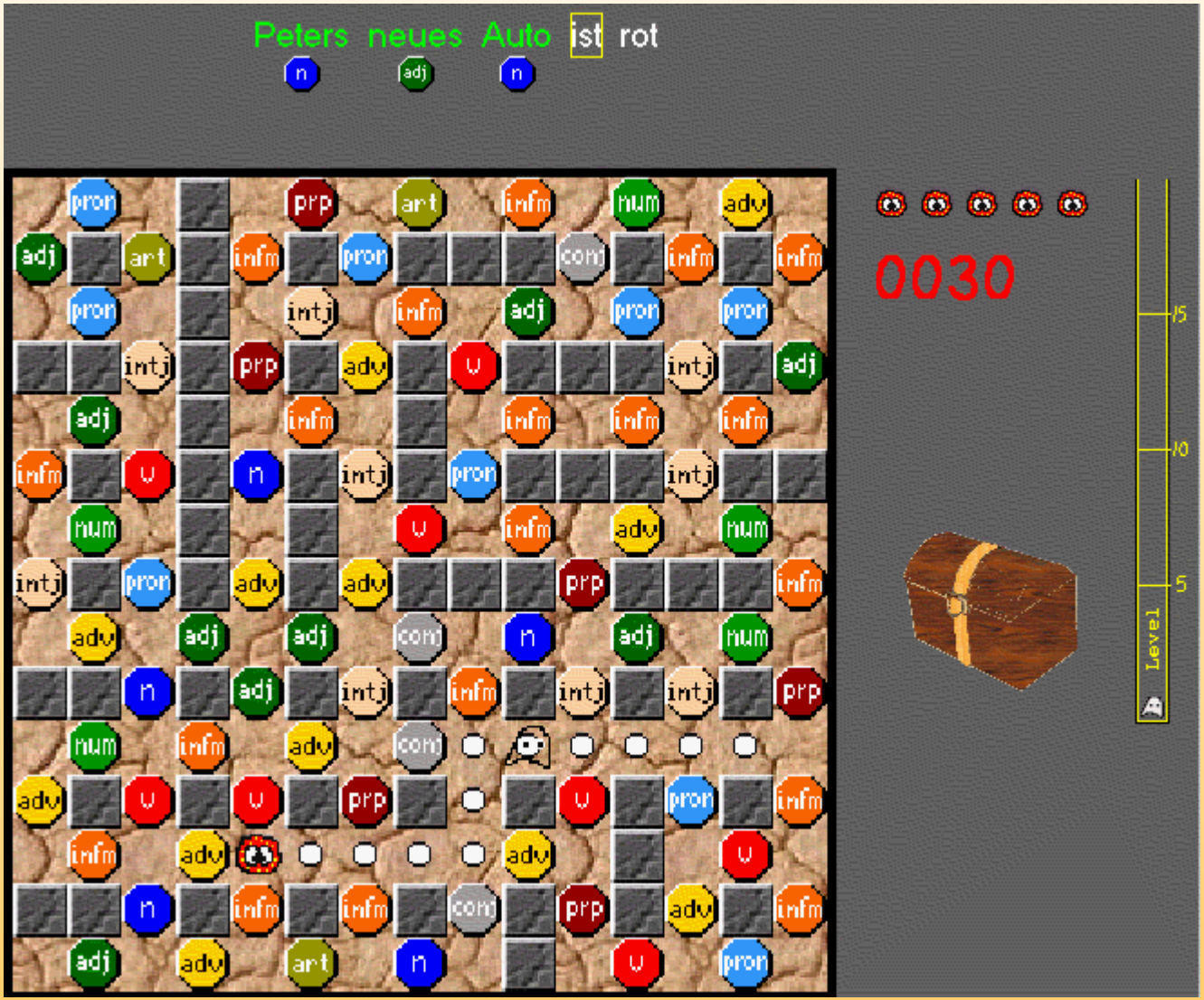
The Paintbox game

ShootingGallery: Hit a noun!



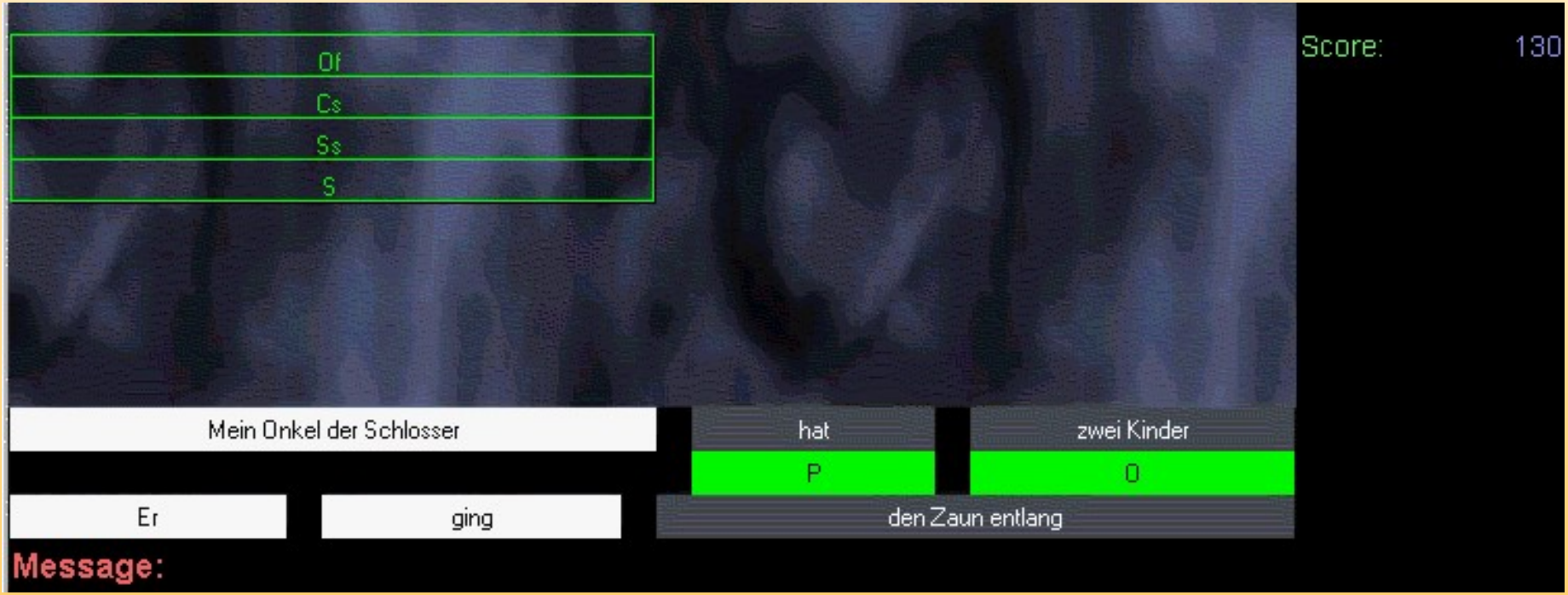
WordFall - Tetris for grammarians

Labyrinth - a word class maze

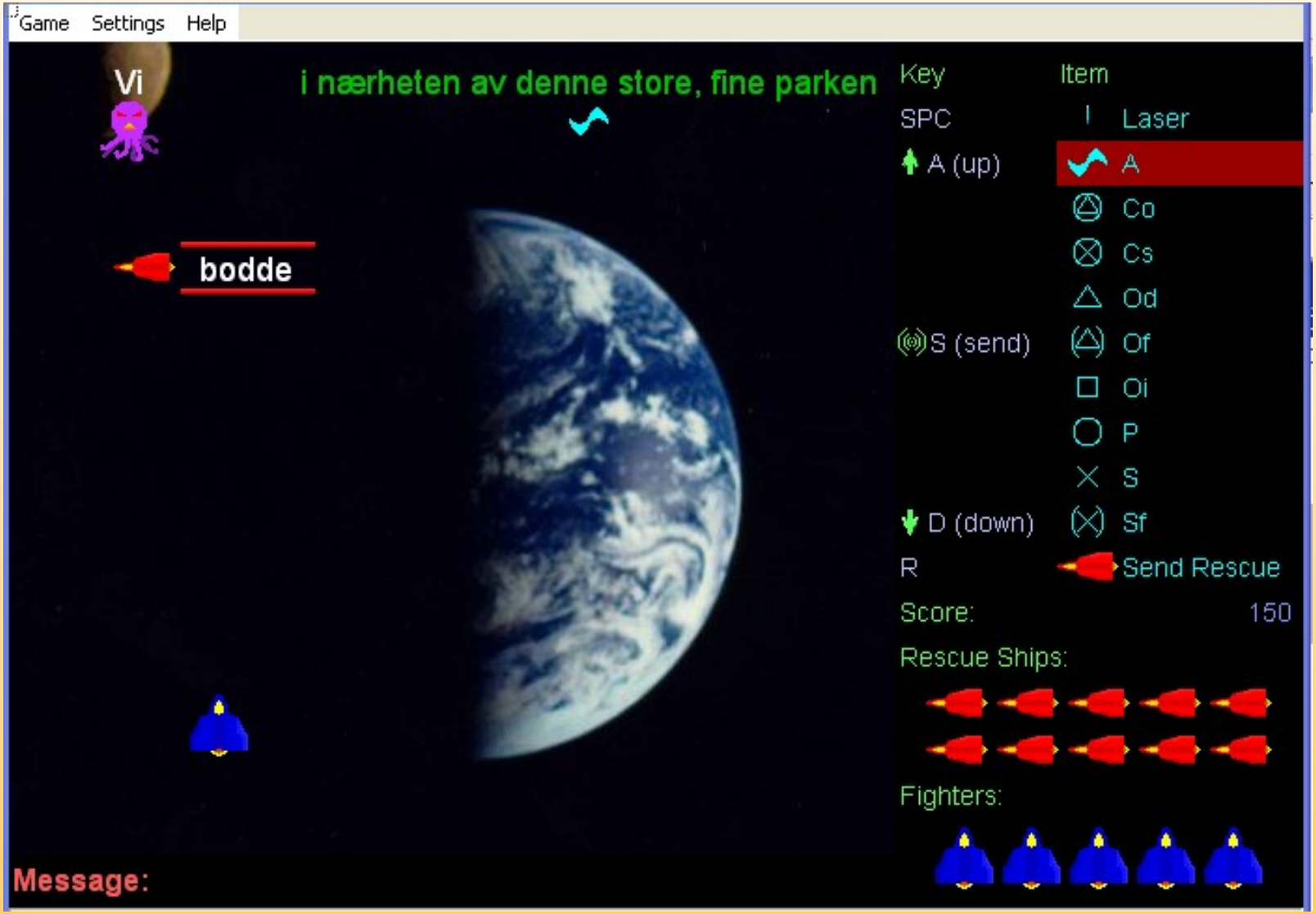


Post office - stamping syntactic function

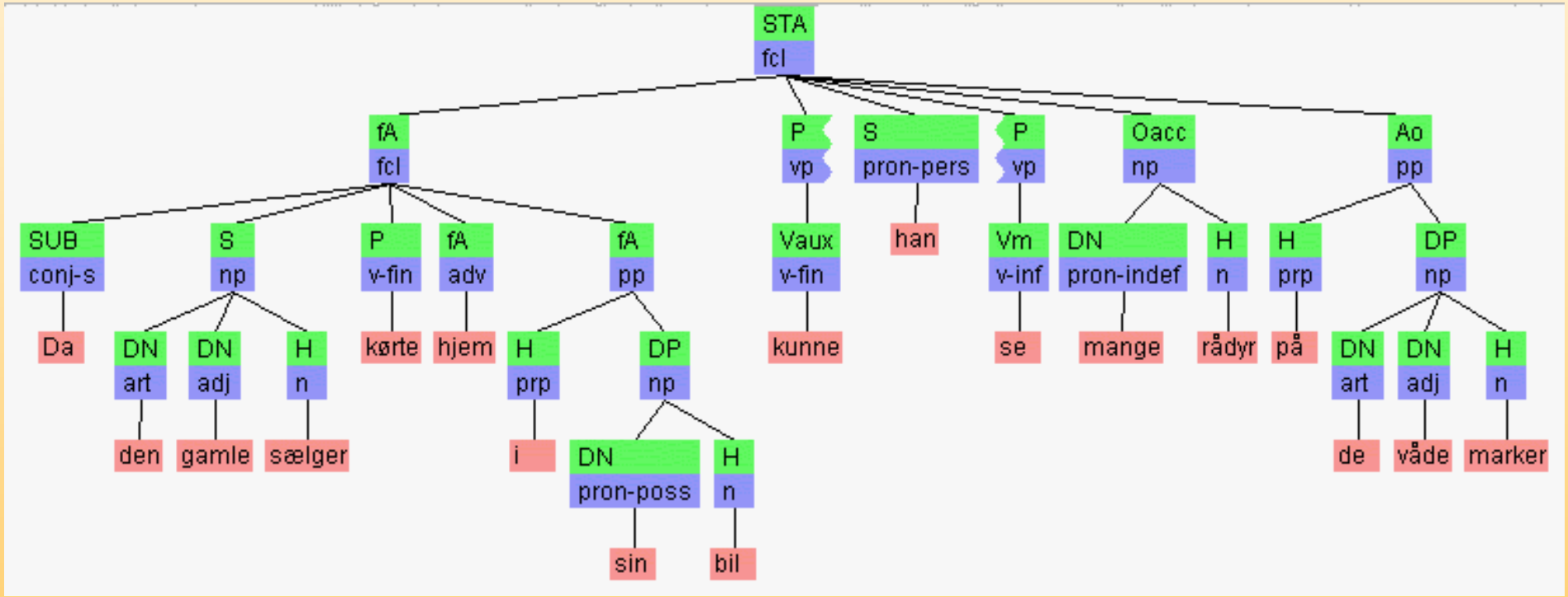
Syntris - syntax brick by brick



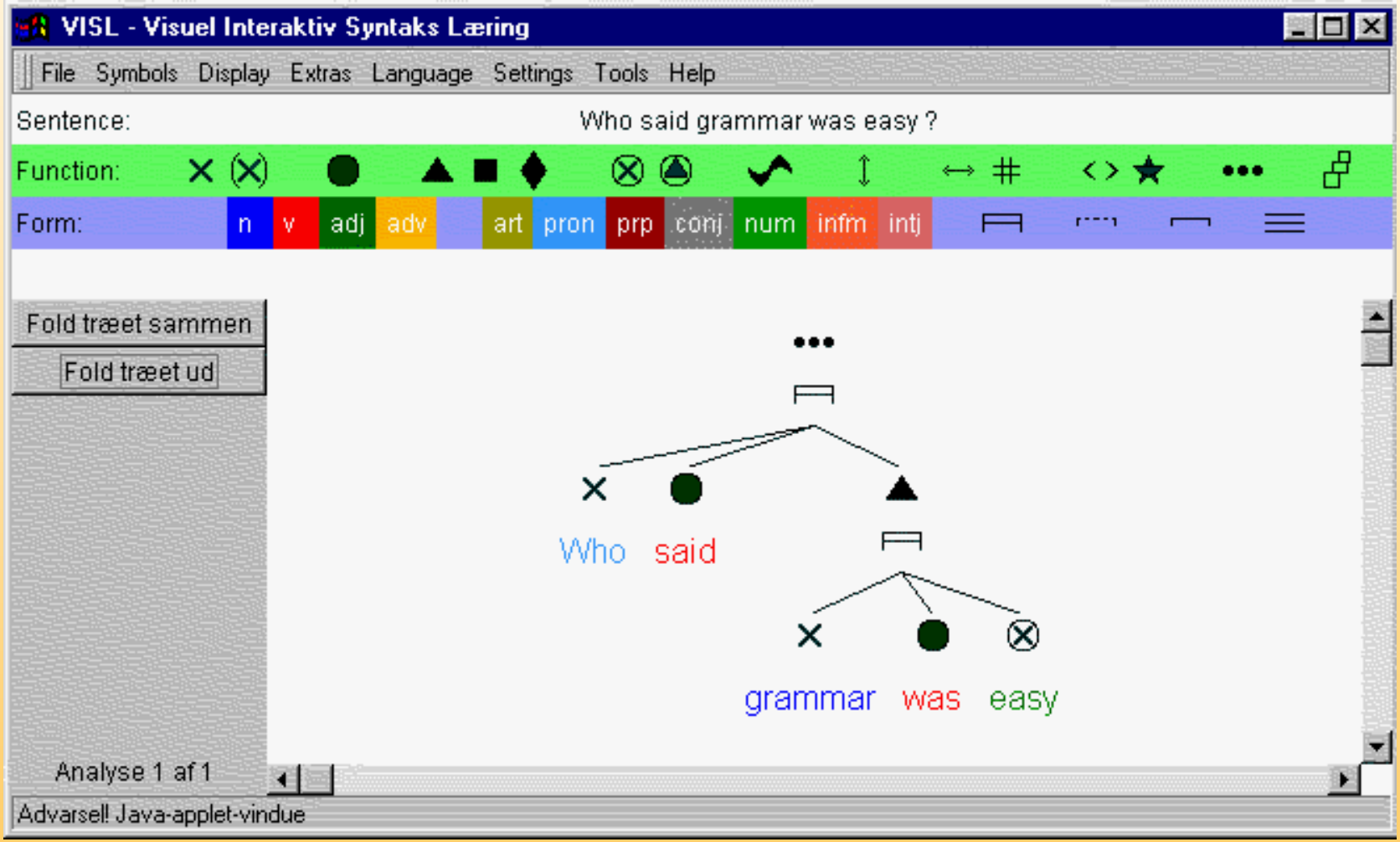
SpaceRescue: Alien syntax



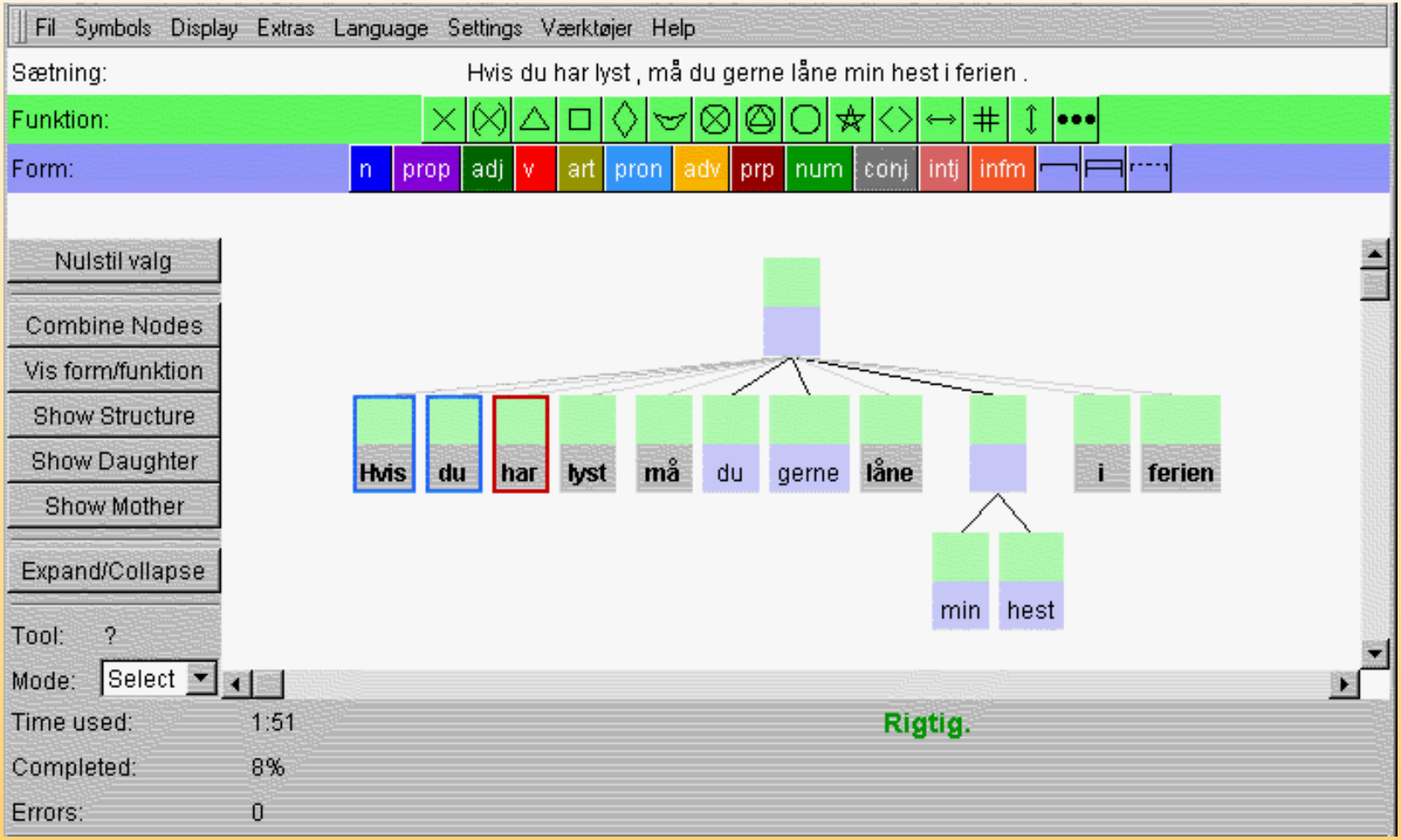
Constituent trees



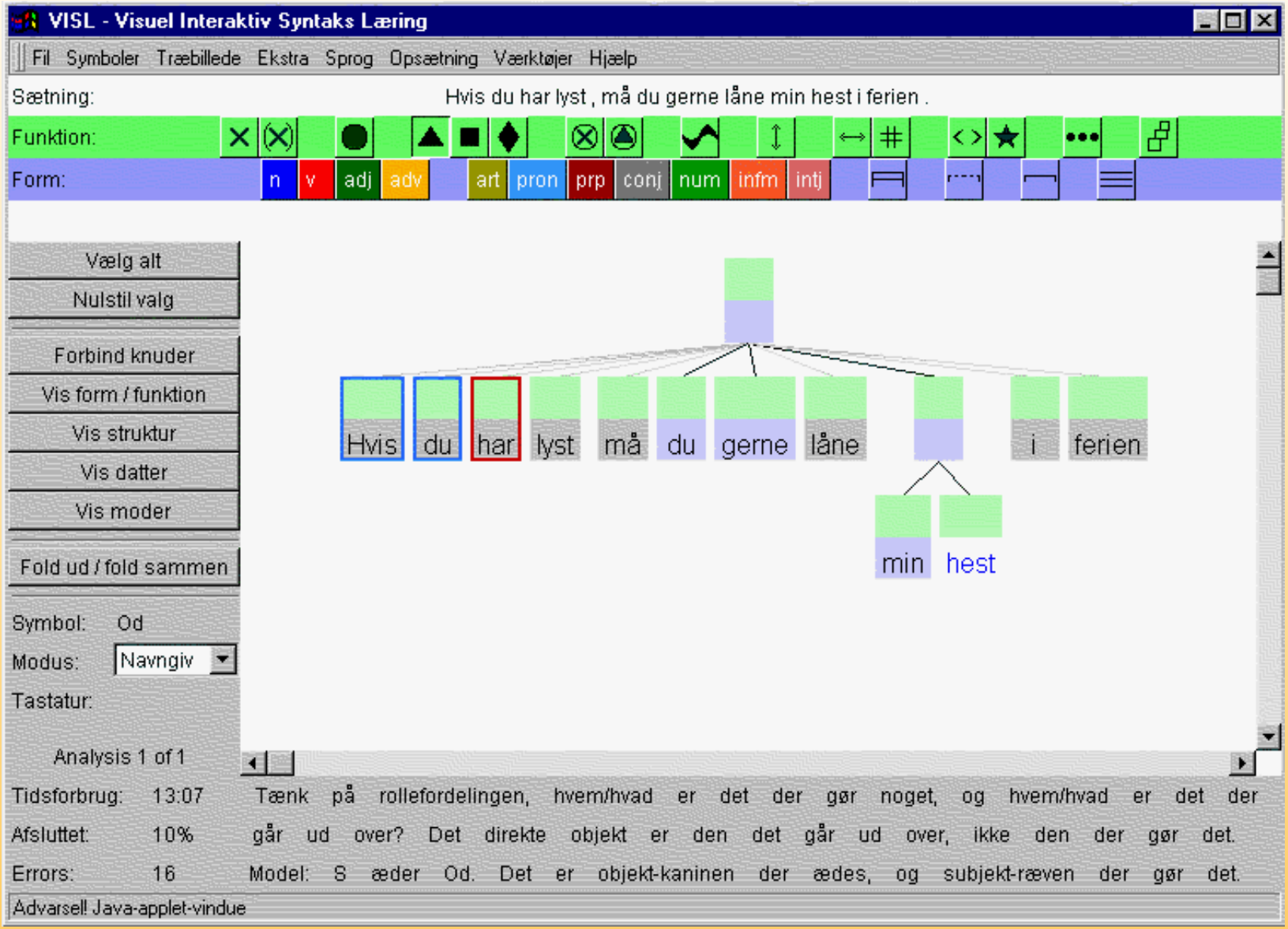
Interactive syntactic trees



BuildTree: Drag & drop constituents



LabelTree: Drag & drop syntactic function



Does it work in real life?

GREI user evaluation

(Oslo University, Kristin Hagen & Janne Bondi Johannessen)

- 3 levels (7th, 8th and 9th grade)
- Use of a VISL group and a control group with traditional grammar teaching.

Before & after testing of VISL and control groups on grammar knowledge after 4 lessons

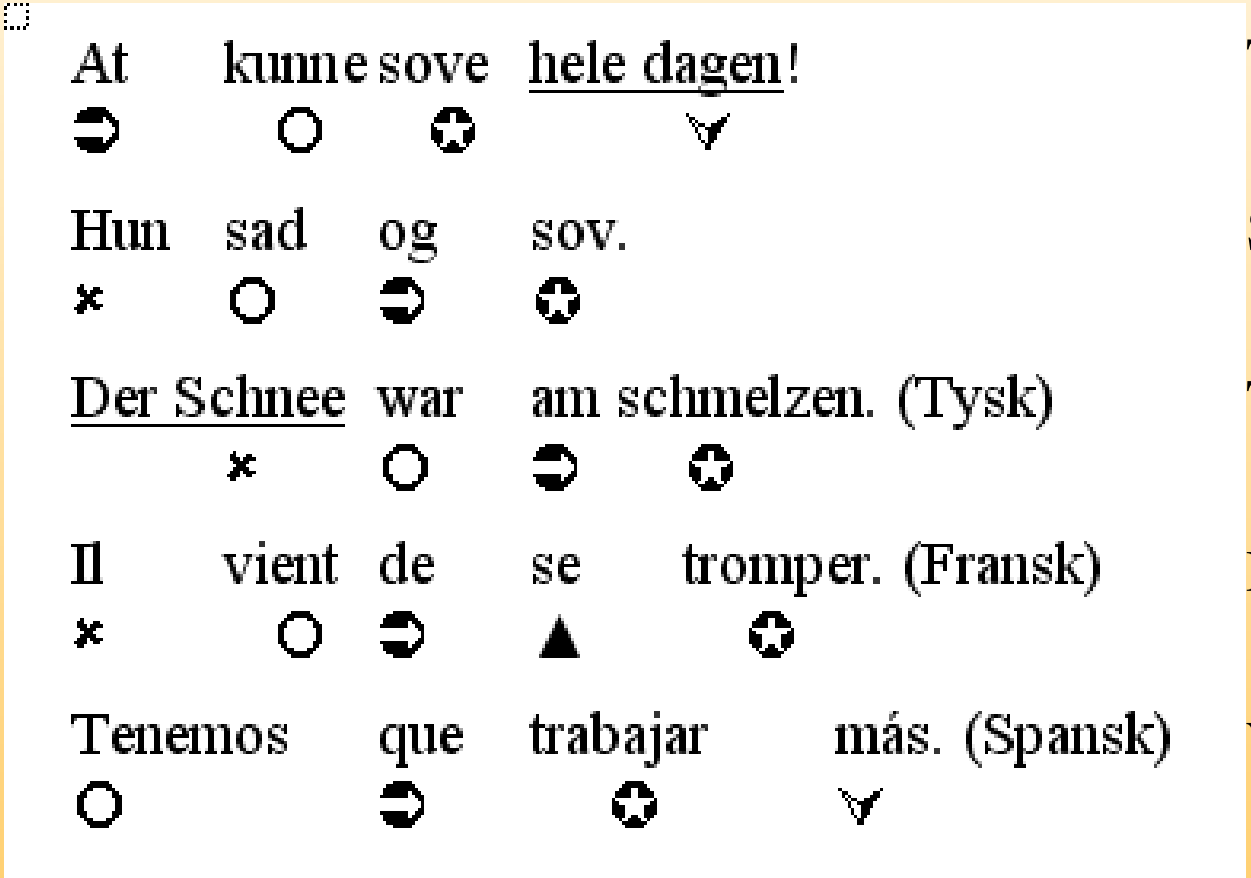
User feed-back

- subjective learning impression: I feel I'm better at grammar now (43% 7th grade, 100% 9th grade)
  - games more fun than syntactic tree-building (100%), but many felt they learned more from the more formal tree-exercise (about 2/5 of 7th grade, 1/4 of 9th grade)

Test results

% improvent in score	Word class	Sentence Analysis	Total
7th grade	1.5% (-3.8%)	17.5% (-2.9%)	11.0% (-3.5%)
8th grade	16.7% (10.5%)	15.2% (6.9%)	15.8% (8.5%)
8th grade	45% (41%)	28.5% (11.3%)	38.6% (26.6%)

Cross-language problems: Infinitive marker



To be able **to** sleep all day

(English default)

She sat (there) **and** slept

(aspect = sleeping)

The snow was melt**ing**

(aspect)

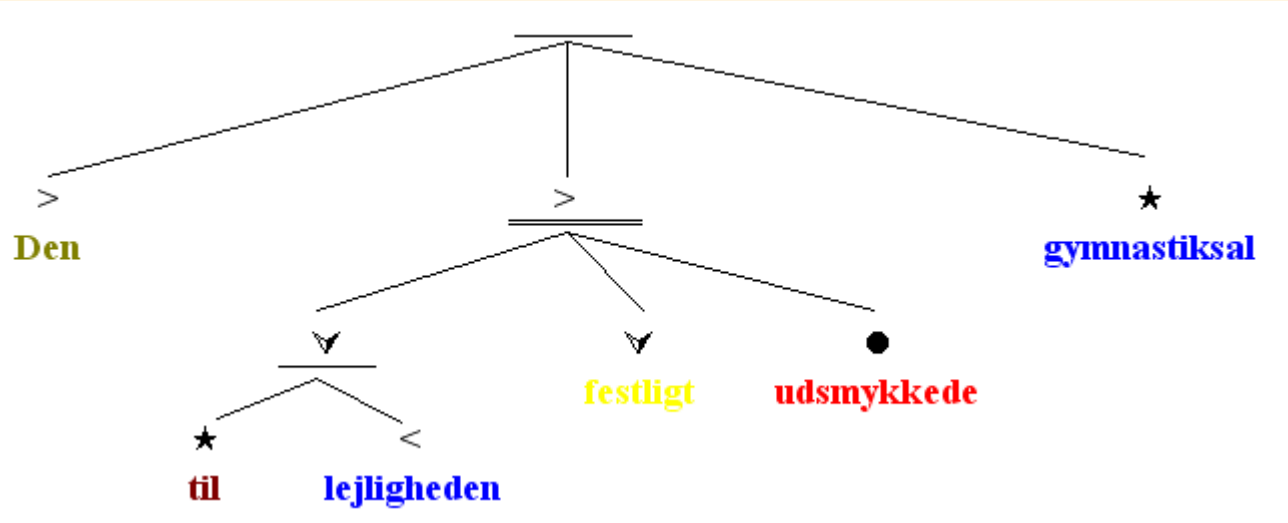
He has **just** made a mistake

(recent past)

We have **to** work

(=“that” we work)

Cross-language problems: participal clauses



English: *Given the fact that ... <u>Once built</u>, the houses ...*

Danish: *Den <u>til lejligheden festligt udsmykkede</u> gymnastiksal*

(The for the occasion lavishly adorned sports hall

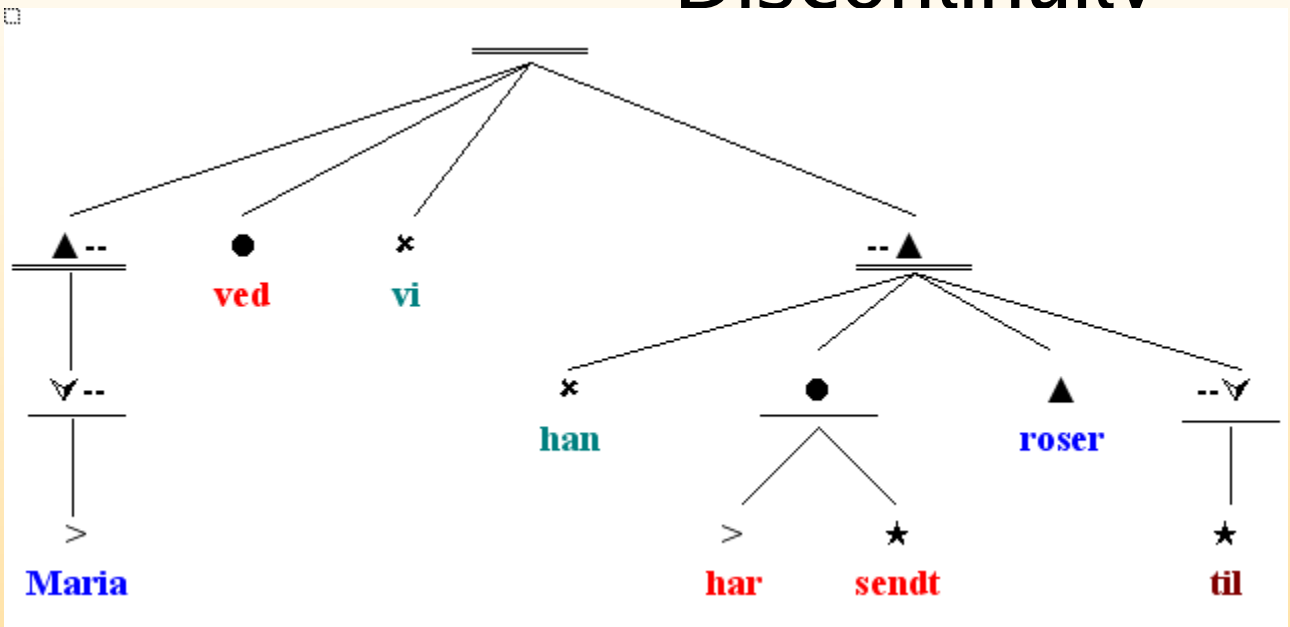
Portuguese: *Feito o trabalho, ... Chegado no aeroporto, ...*

(Finished the work,... Arrived at the airport, ...)

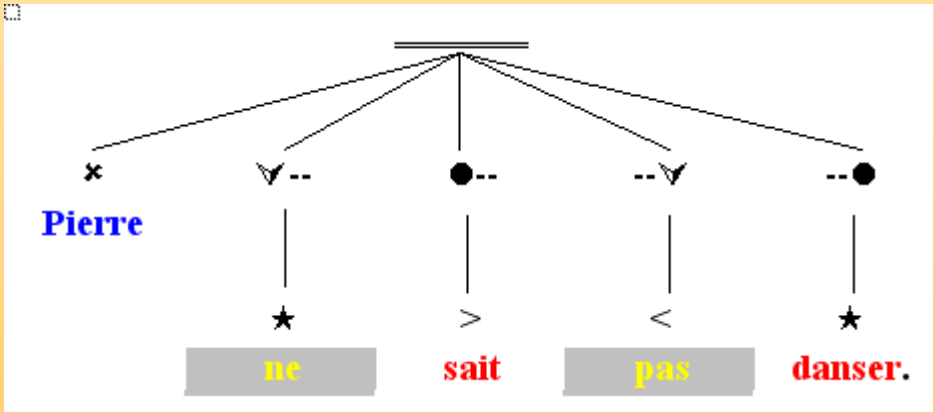
German: *Der <u>vom Rat genehmigte</u> Zuschuss*

(The subsidies conceded by the Council)

Cross-language problems: Discontinuity



Marta know we he has sent roses to



Pierre not can not dance

VISL source notation

VISL lite vertical tree (non-graphical notation, filtered)	**VISL vertical tree** (non-graphical notation, incl. morphology)
UTT:cl	STA:fcl
S:prop VISL	S:prop("VISL") VISL
P:v er	P:v-fin("være",pr,akt) er
Cs:g	Cs:np
=D:art et	=DN:art("en",neu,sg,idf) et
=H:n forskningsprojekt	=H:n("forskningsprojekt",neu,sg,idf,nom) forskningsprojekt
=D:cl	=DN:fcl
==S:pron der	==S:pron-rel("der",nG,nN,nom) der
==P:v involverer	==P:v-fin("involvere",pr,akt) involverer
==Od:g	==Od:np
===D:pron mange	===DN:pron-indef("mange",nG,pl,nom) mange
===D:adj forskellige	===DN:adj("forskellig",nG,pl,nD,nom) forskellige
===H:n sprog	===H:n("sprog",neu,pl,idf,nom) sprog

CG source notation (function/dependency)

VISL	[VISL]	<heur><*>	**PROP** NOM	@SUBJ>	#1->2
er	[være]	<vk>	**V** PR AKT	@FMV	#2->0
et	[en]		**ART** NEU S IDF	@>N	#3->4
forskningsprojekt	[forskningsprojekt]		**N** NEU S IDF NOM	@<SC	#4->2
,					#5->0
der	[der]	<rel>	**INDP** nG nN NOM	@SUBJ>	#6->7
involverer	[involvere]	<vt>	**V** PR AKT &MV	@FS-N<	#7->4
mange	[mange]	<quant>	**DET** nG P NOM	@>N	#8->10
forskellige	[forskellig]		**ADJ** nG P nD NOM	@>N	#9->10
sprog	[sprog]		**N** NEU P IDF NOM	@<ACC	#10->7
.					#11->0

Supported xml-formats

- TIGER-xml (constituents)
- TIGER-xml (dependency)
- MALT-xml
- VISL data file markers:
  pedagogical topic and chaptering attributes
  for dynamic html-layout

The advantage 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”

The Portuguese example

- Portuguese object pronouns need an “attractor” (negation, subject) in order to allow pre-verbal position
- More so in Portugal than in Brazil or Mozambique
- Diachronic fluctuation, sociolect / speaker status
- Introspection gives normative results
- Corpus gives true(er) results (NURC, Tycho Brahe, Folha vs. Público ....)

How to enrich a corpus

- Meta-information: Source, time-stamp etc.
- Grammatical annotation: Part of speech (PoS), inflexion, syntactic function, syntactic structure, semantics ...
- Manual vs. automatical annotation

e.g. Korpus90 and Korpus2000

- mixed text, ca. 20 (28) mill. ord each
- sentence-randomized “quote” corpus
- compiled by DSL (www.dsl.dk)
- grammatically annotated by VISL (visl.sdu.dk)
  - a) automatically with the DanGram parser
  - b) 1% manually revised (Arboretum treebank)

How to annotate

- All annotation is theory dependent, but some schemes less so than others. The higher the annotation level, the more theory dependent
- double role of corpora: (a) as goal, (b) as (gold-standard annotated) data for machine learning: rule-based systems for boot-strapping
- PoS (tagging): needs a lexicon ("real" or corpus-based)
  (a) probabilistic: HMM-base line, DTT, TnT, Brill etc., F ca. 97+%
  (b) rule-based:
  --- Disambiguation as a "side-effect" of syntax (PSG etc.)
  --- Disambiguation as primary method (CG), F 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)

Constraint Grammar

- A methodological rather than descriptive paradigm (Karlsson 1995) Token-based assignment and contextual disambiguation of tag-encoded grammatical information
- Grammars need lexicon/analyzer-based input and consist of thousands of MAP, SUBSTITUTE, REMOVE and SELECT rules.
  - e.g. REMOVE (@<SUBJ) (NOT 0 N-HUM) (*-1 V-HUM BARRIER NON-PRE-N LINK 0 AKT) ;
  - SELECT (ADJ + MS) (-1C ART + MS) (*2C NMS BARRIER NON-ATTR OR (F) OR (P)) ;
- The VISL project (SDU) uses Constraint Grammar parsers to add form and function tags to word tokens in corpora or running text
- Form: e.g. N = noun, P = plural, GEN = genitive
- Syntactic function: e.g. @SUBJ = subject, @ACC = direct object
- Syntactic form: e.g. dependency markers (@SUBJ>, @<SUBJ), numbered dependency (e.g. #5->3) or secondary constituent trees

A dependency grammar for CG input

(c1) @FS-@N< -> (¤NPHEAD, N.*@N<)
IF (L) TRANS:(@SUBJ>,@F-SUBJ>,@S-SUBJ>)

(c2) @ADVL> -> (<mv>)
IF (R) BARRIER (@SUBJ>,@F-SUBJ>,@S-SUBJ>

(c3) <np-close> -> (DET)
IF (L) HEADCHILD=(@>N)

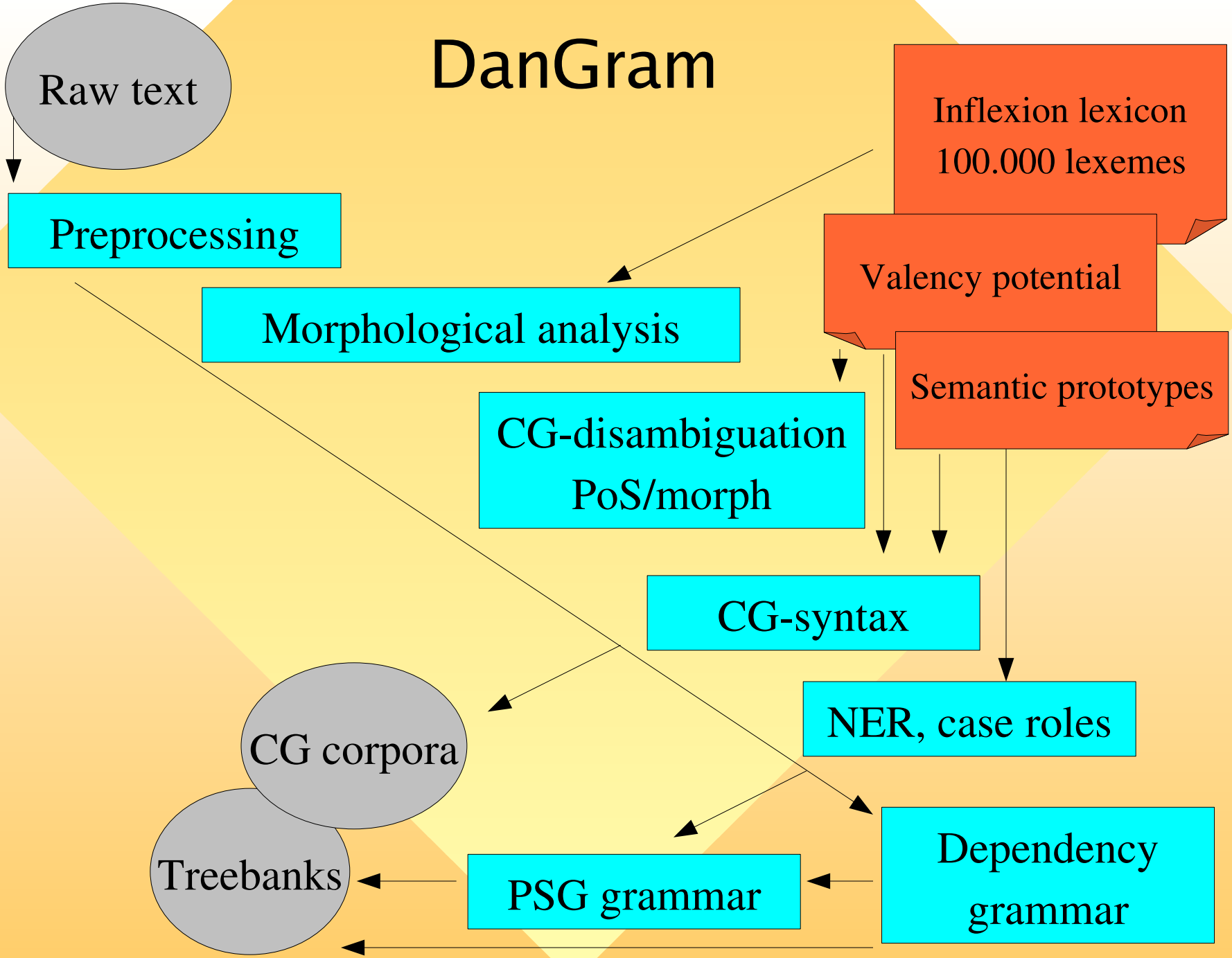
(c4) @N< -> (N,PROP,PERS,INDP,¤NPHEAD)
IF (L) NOTHEAD=(<clb>) NOTTARGET=(@FS-@N<)

The grammar respects head-uniqueness, and tries to avoid circularities. It allows forced and inverted attachments, as well as set definitions.

Evaluation of the Danish system (TLT05)

1437 words 1663 tokens	*errors*	*accuracy (words, not tokens, out of all)*
Part of speech *- on raw text*	10	99.4 %
Syntactic function (edge label) *- on raw text*	73	95 %
Dependency (attachment) *- on raw text*	102	93 %
Dependency *- on function-corrected input*	20	98.7 %

DanGram


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

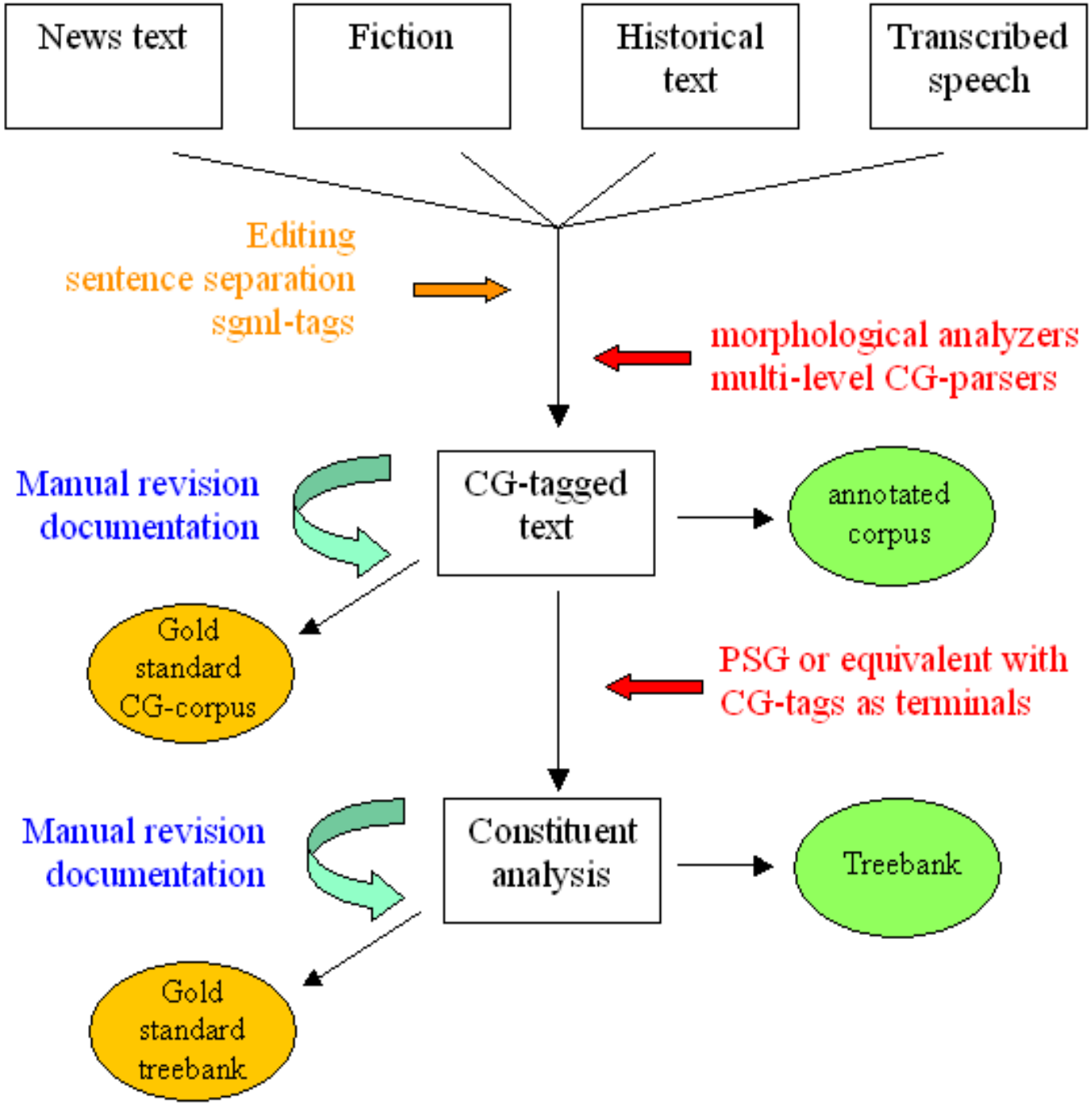
Cg-results for Danish: PoS

Class	recall	precision	F-score	Class	recall	precision	F-score
N	99.5	99.1	99.2	**ART**	99.3	99.3	99.3
PROP	100	100	100	**DET**	97.1	98.5	97.7
V PR	99.2	99.2	99.2	**PERS**	99.4	99.4	99.3
V IMPF	100	97.2	98.8	**INDP**	98.2	100	99.2
V INF	98.1	99.0	98.5	**NUM**	100	100	100
V PCP1	100	100	100	**ADJ**	96.8	94.4	95.5
V PCP2	94.9	97.4	96.1	**ADV**	95.8	98.0	96.8
INFM	100	100	100	**PRP**	99.4	99.1	99.2

CG-result for Danish: Syntactic function

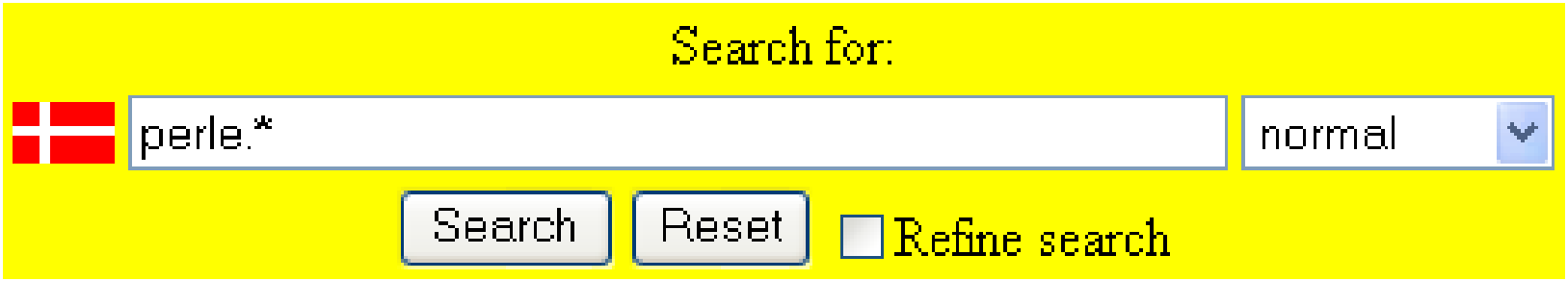
Class	recall	precision	F-score	Class	recall	precision	F-score
@SUBJ>	96.7	95.2	95.9	@>N	97.3	98.2	97.7
@<SUBJ	90.1	96.8	93.3	@N<	90.9	96.1	93.4
@F-SUBJ>	86.6	86.6	86.6	@APP*	100	87.5	93.3
@F-<SUBJ	100	100	100	@N<PRED	100	80.0	88.8
@<ACC	94.6	95.3	94.9	@>A	88.6	95.9	92.1
@ACC>*	88.8	88.8	88.8	@A<	89.4	94.4	91.8
@<DAT*	100	75.0	85.7	@P<	98.1	98.1	98.1
@<PIV	93.5	87.8	90.5	@FS-<SUBJ*	77.7	77.7	77.7
@<SC	92.0	84.3	87.9	@FS-<ACC	100	72.7	84.1
@<OC*	83.3	100	90.8	@FS-ACC>	100	91.6	95.6
@<SA	83.3	86.9	85.0	@FS-<ADVL	90.3	96.5	93.2
@<OA*	100	75.0	86.7	@FS-ADVL>	84.6	78.5	81.4
@<ADVL	93.2	90.6	91.8	@FS-P<	90.9	100	95.2
@ADVL>	96.9	93.2	95.0	@ICL-<SUBJ*	100	100	100
@KOMP<*	100	100	100	@ICL-P<	96.1	100	98.0

Corpus annotation



The interface

Simple text searches: e.g. Composita / affixes



... de las sociedades occidentales reside en la **hipertrofia** de el individualismo jurídico

Eficacia e **hiperreglamentación** no van parejas .

... sufre una crisis estructural y mercados rígidos e **hiperregulados** .

... de satélites , de antenas , de ordenadores hiperpoderosos , utilizando ...

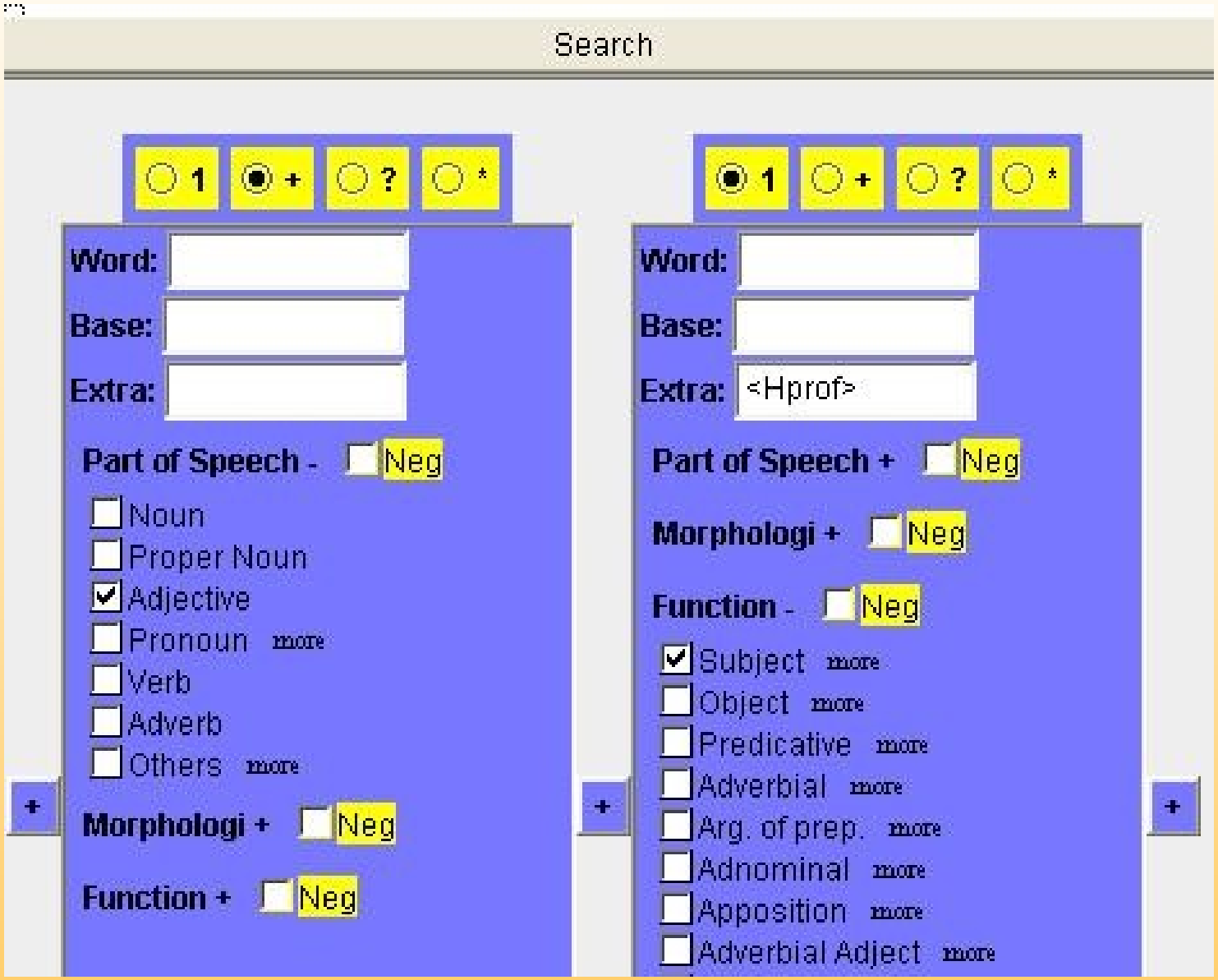
... éste a la existencia de estas formas de trabajo **hiperflexibilizadas** ?

... a el cabo , legitimar a estos precursores de la **hiperflexiblidad** .

... el mito de que se puede ser " guapos , potentes e **hipercativos** " sin esfuerzo .

... traslados de empresas , desertización rural , **hiperconcentración** urbana ...

Menu-based searches



Statistical tools

corpuseye Help Troubleshooting Grammatical information Taglist

sort freq rel

By: Left Context / Right Context / Left Edge / Right Edge

Offset 0

Freq items 100

Refine search

New search

Searched for: [pos="((.*)?ADJ(.*)?)"] [extra="Azo"]
In corpus: DAN_C2000
Found 613 results (613).
1 - 50 next

INFO	Den nu	**fire-årige hanbjørn** er kendt
INFO	kommer dundrende ned fra bakkedraget med en	**gigantisk hjort** tøvende i bag
INFO	Et spil der appellerer til den	**politiske ræv** , fordi det kan
INFO	havde juryen indført en ekstra jurypris med	**tilhørende sølvbjørn** til den
INFO	Hans blik minder ikke længere en	**skræmt hjort** fanget i forlygt
INFO	Hvis der i_fjor var antydning af tvivl om den	**amerikanske tigers** holdbarhed
INFO	og engagement på konferencen - ikke kan få den	**russiske bjørn** til at gungre
INFO	Gys , gru og store	**stygge ulv** .
INFO	mest politiske bestilling - og Danmarks mest	**berømte løve** på en sokkel .
INFO	negre " - styrken ved at være i kontakt med den	**indre abe** er i_hvert_fald til
INFO	Og jeg ved , hvor der er nogle	**store frøer** nede i skoven , s
INFO	sig rive med fra starten og prøve at følge med de	**unge løver** , der vil køre sta
INFO	- ANDi ser almindelig ud , mens et_par af de	**dødfødte aber** lyste fluoresce
INFO	Imidlertid er en	**grøn marekat** en sand gourmet

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
- **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**, transscribed Brazilian speech, 100K words
- **Tycho Brahe**, historical Portuguese, 50K words

Treebanks

- **Floresta Sintá(c)tica**, European Portuguese, 1M words (200K revised)
- **Arboretum**, Danish, 200-400K words revised

The case for treebanks

- A treebank is a corpus annotated with full syntactic structure, attaching tokens to each other (dependency grammar) or to interconnected non-terminal nodes (constituent grammar)
- Treebanks contain more syntactic detail than tagged corpora
- Treebanks allow to train or evaluate automatic systems of analysis
- Treebanks allow searches for complex units and their relations, rather than individual tokens or their features. For instance, the sequence of NPs with certain functions can be queried directly, or conditioned on their being daughters of an embedded clause (subclause).
- Treebanks exist for a large number of languages (cp. CoNLL-X shared task), e.g. Negra/TIGER (German), Penn (English), Mamba (Swedish), Cast3LB (Spanish) ....
- The largest VISL treebank is the double-format Arboretum treebank for Danish, annotated in both dependency and constituent grammar

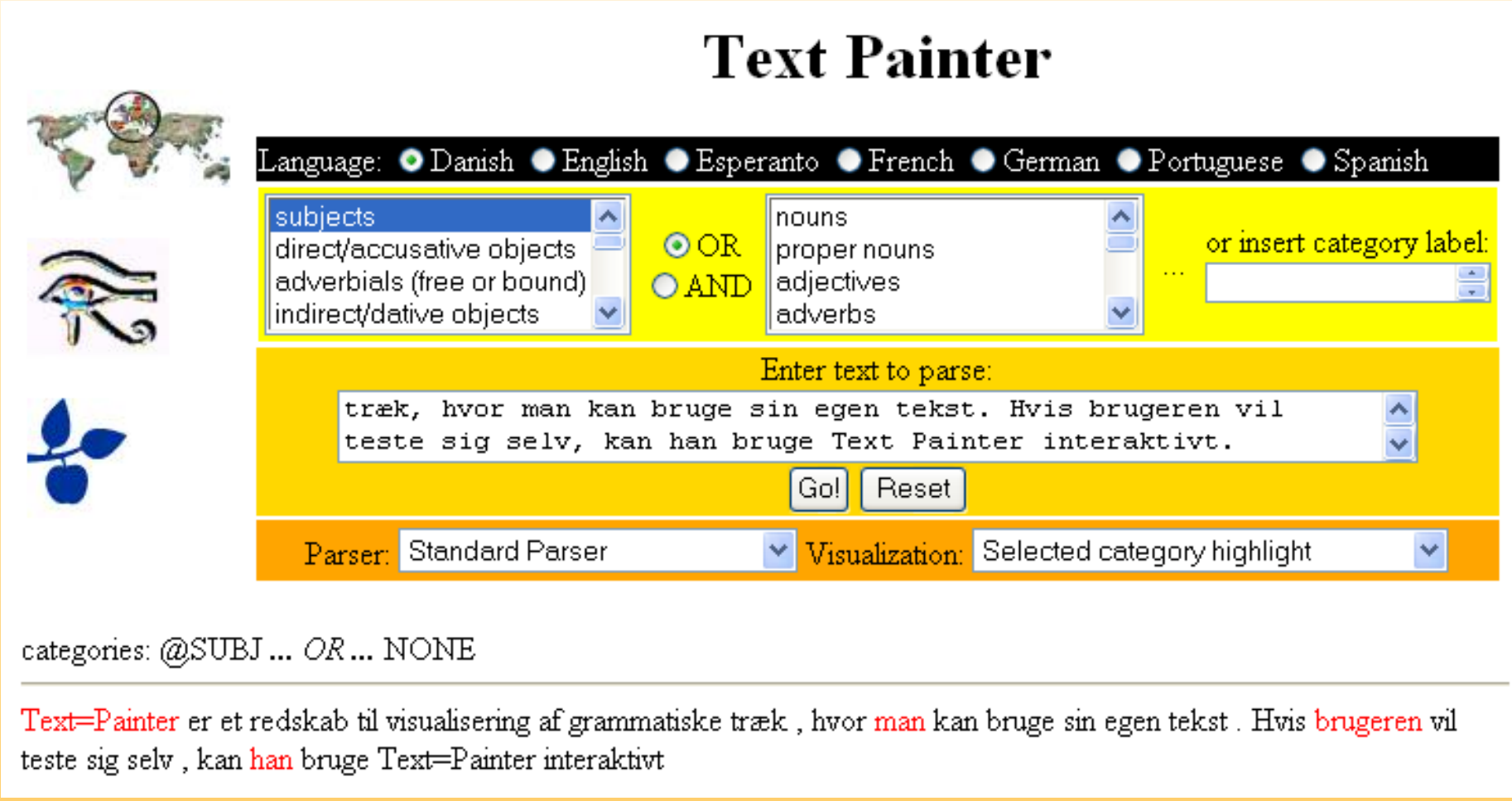
Google as a corpus

- Advantages
  - Much larger than any existing corpus
  - Very accessible
  - Contains data close to spoken language (chats, blogs, discussion fora)
- Disadvantages
  - Can't search for lemma, PoS or syntactic function
  - Difficult to control genre, language level, diachronics
  - Frequencies are not accurate (doubles etc.)
  - No subsorting/statistics for adjacent tokens
  - Results are harder to sift through (no concordance or alphabetical sorting)

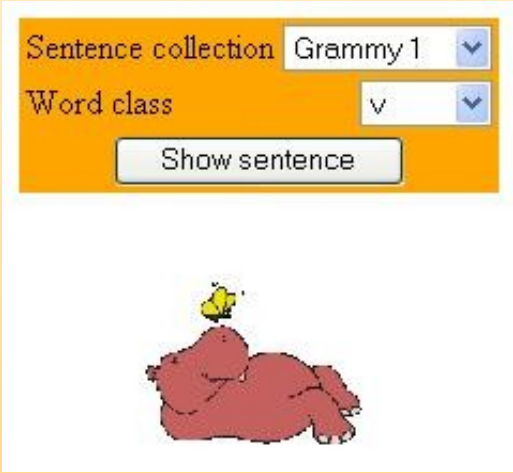
Nevertheless

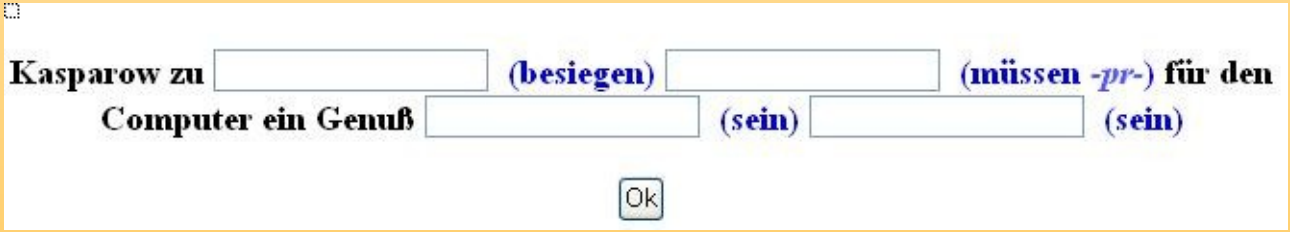
- Qualitative vs. Quantitative (e.g. language awareness)
  - Find examples (at all)
  - Check variation (e.g. Official vs. factual usage)
  - Regional usage (site:/domain)
- webcorp: Searching the internet as a corpus, slow but nice: http://www.webcorp.org.uk/
- web-conc: Concordancing with the whole internet as a corpus. http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi
- The internet as a monitor corpus: http://www.it.usyd.edu.au/~vinci/webcorpus.html
- Robb T. (2003) "Google as a Quick 'n Dirty Corpus Tool": http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html

Integrating live NLP and language awareness teaching

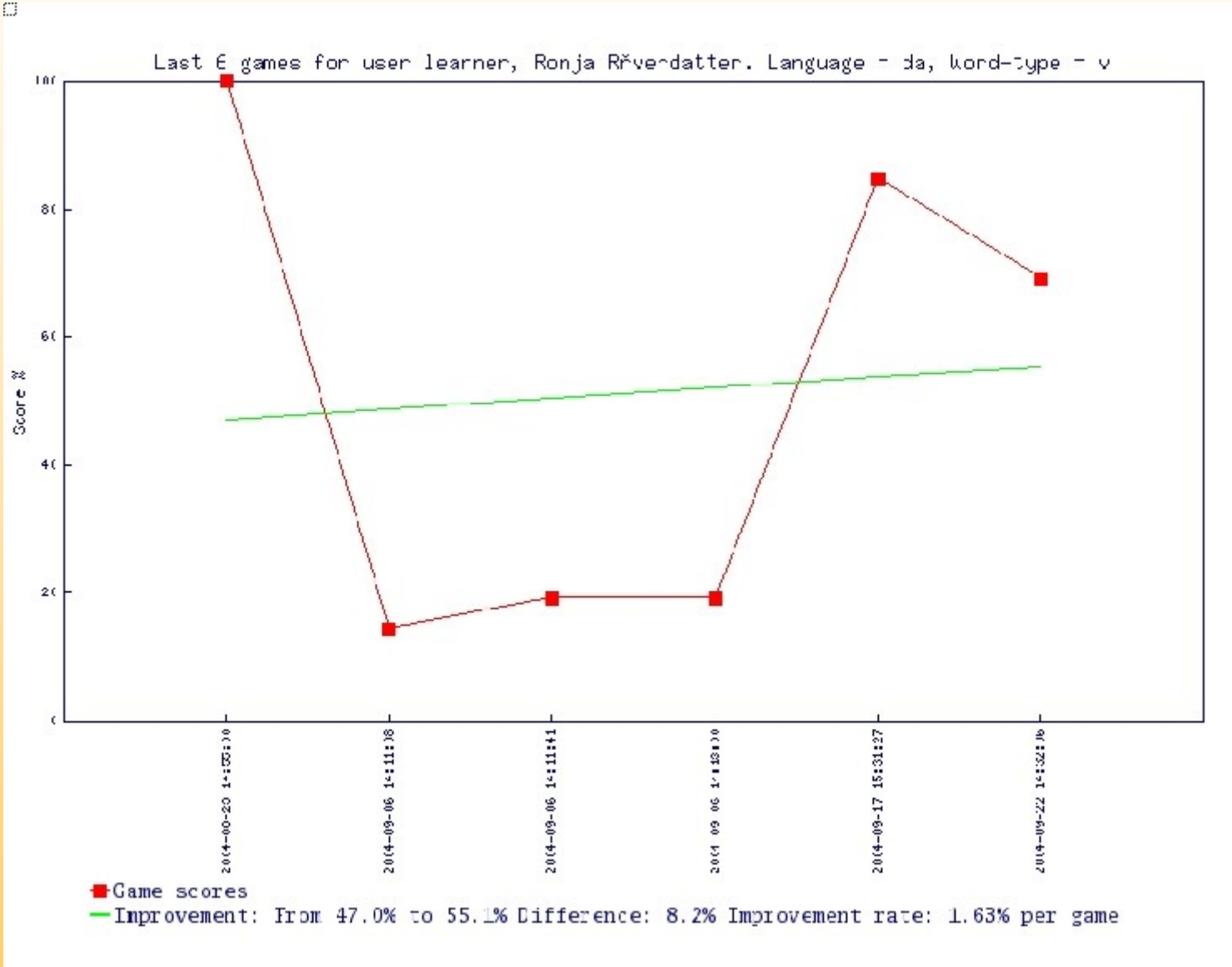


KillerFiller: Towards evaluation





Performance statistics



VISL

http://visl.sdu.dk

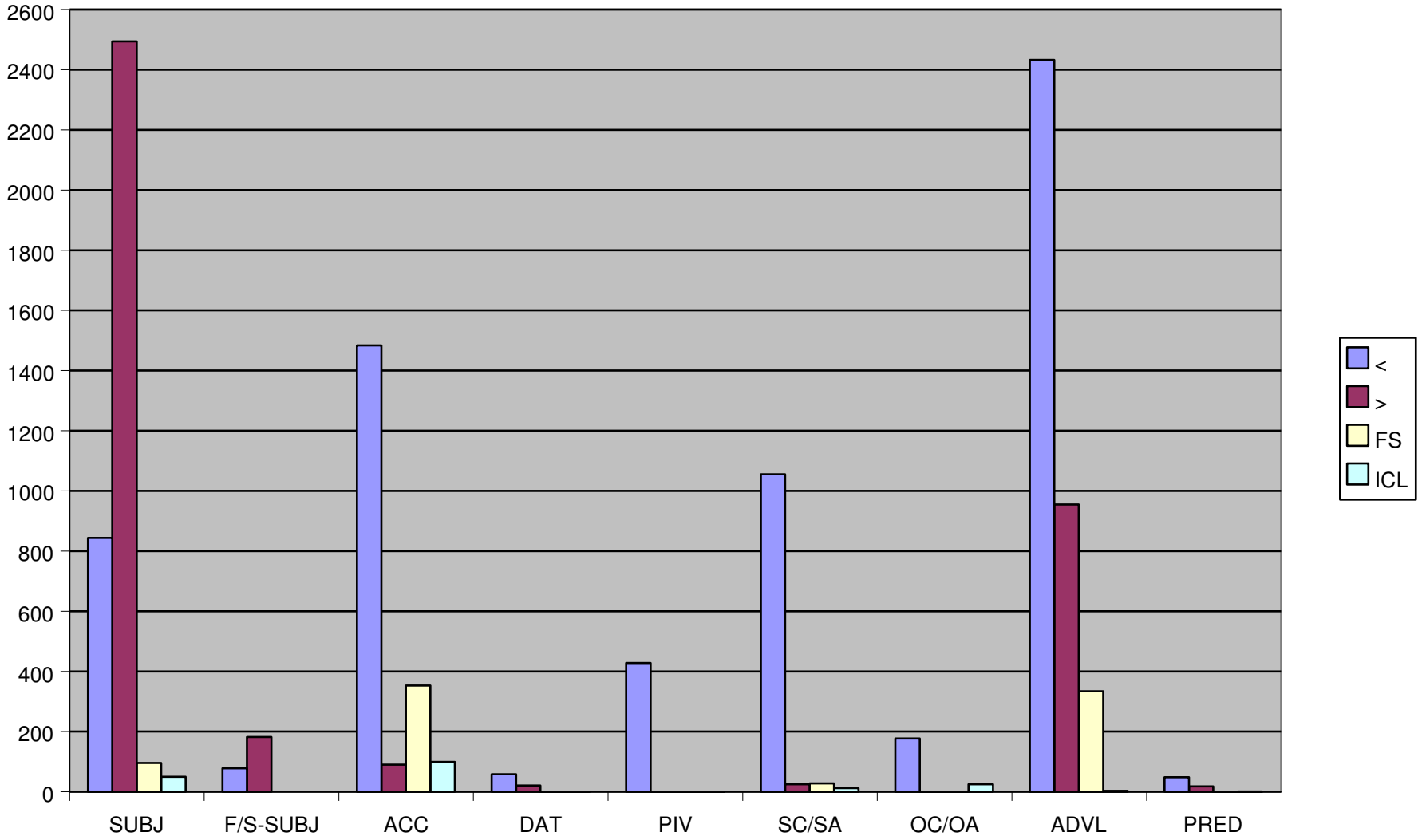
Eckhard Bick, lineb@hum.au.dk

****

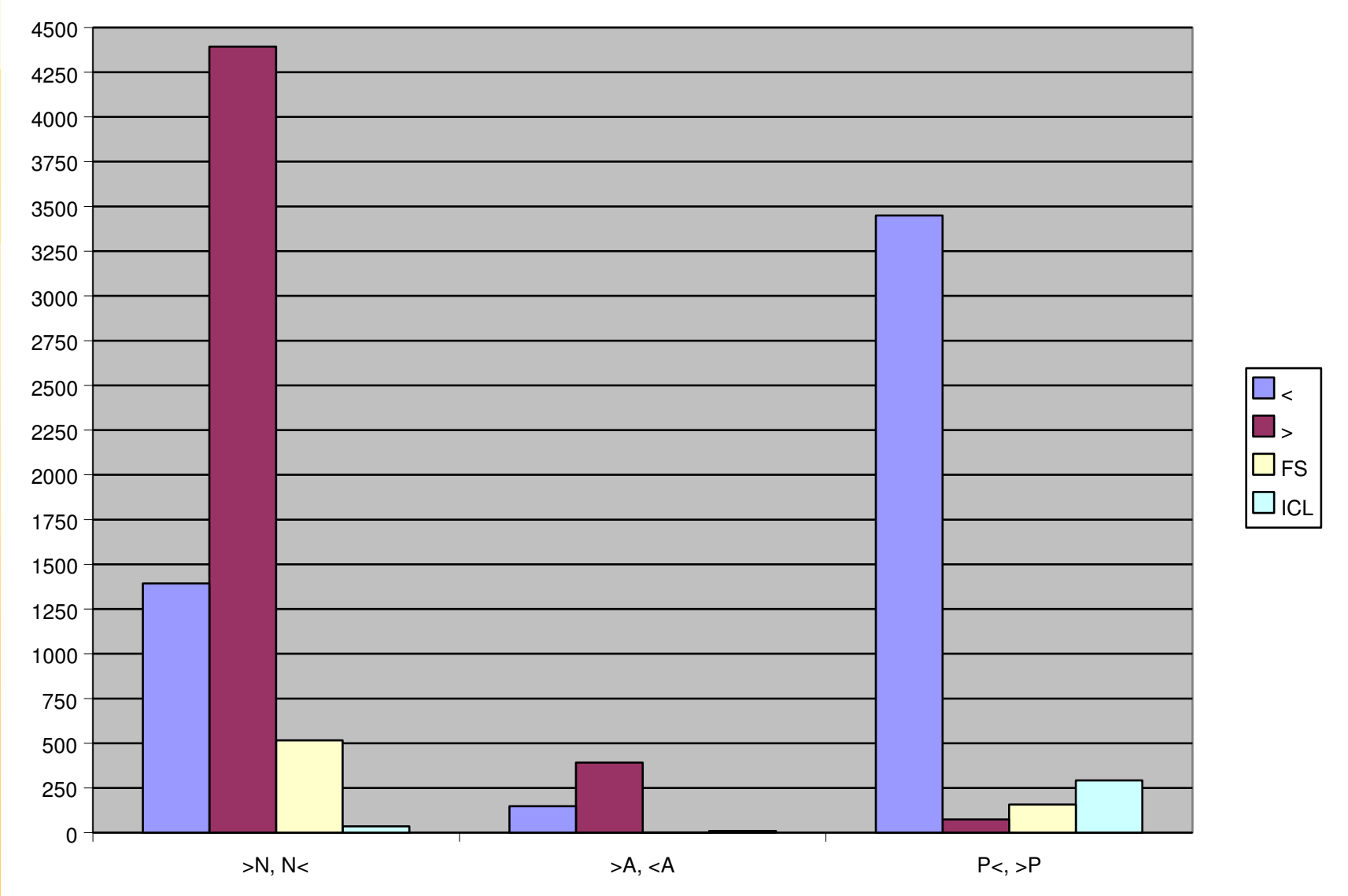
The most common syntactic categories

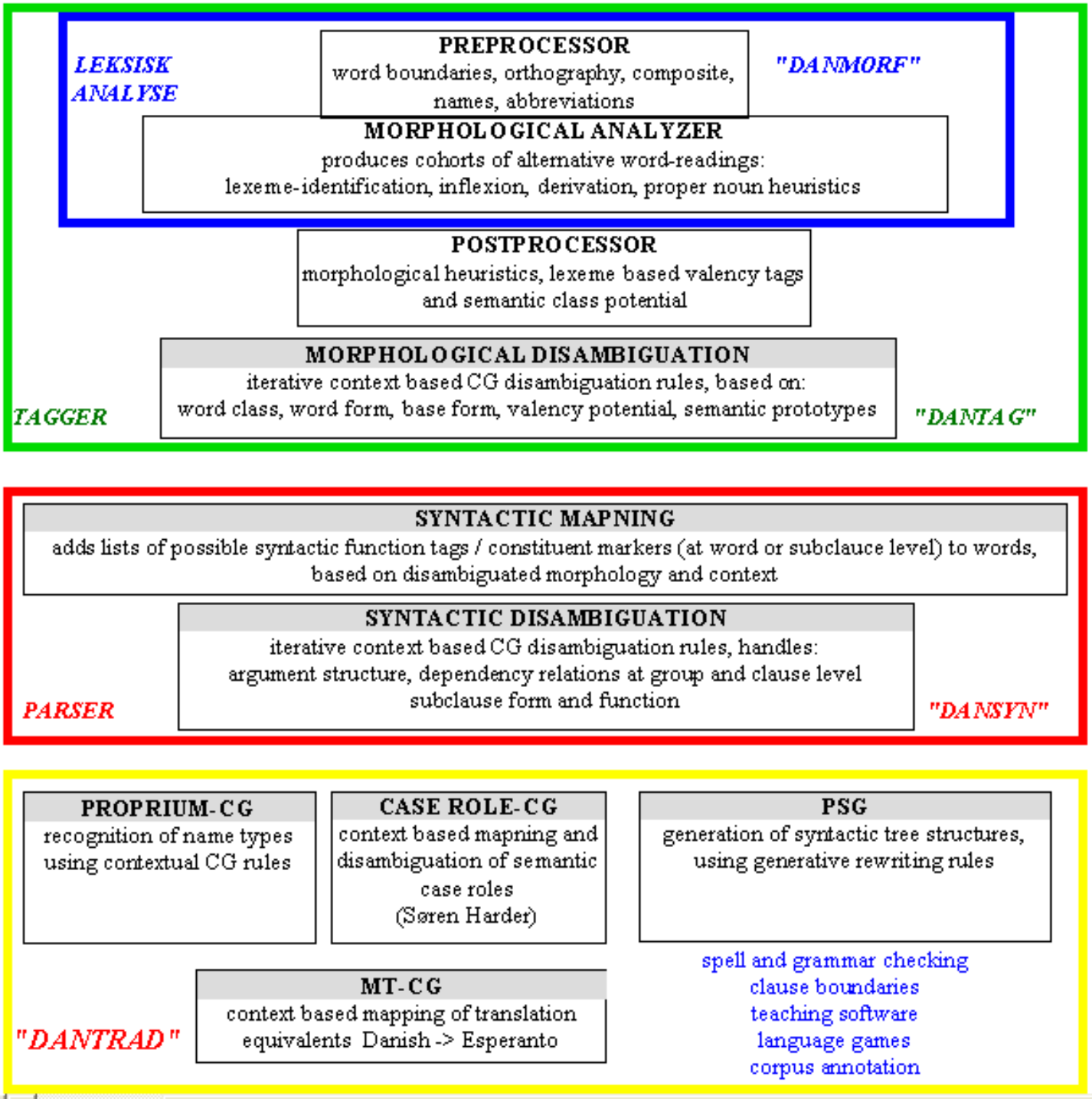
@SUBJ	subject	**@ADVL**	free (adjunct) adverbial
@ACC	direct (accusative) object	**@PRED**	free (adjunct) predicative
@DAT	indirect (dative) object	**@APP**	apposition
@PIV	prepositional object	**@>N**	prenominal dependent
@SC	subject complement	**@N<**	postnominal dependent
@OC	object complement	**@>A**	adverbial pre-dependent
@SA	subject related adverbial argument	**@A<**	adverbial post-dependent
@OA	object related adverbial argument	**@P<**	argument of preposition
@MV	main verb	**@INFM**	infinitive marker
@AUX	auxiliary	**@VOK**	vocative

Clause level dependents, left/right distribution in Korpus90/2000



Modifier position, distribution in Korpus90/2000




LEKSISK ANALYSE
PREPROCESSOR
word boundaries, orthography, composite, names, abbreviations
"DANMORF"
MORPHOLOGICAL ANALYZER
produces cohorts of alternative word-readings:
lexeme-identification, inflexion, derivation, proper noun heuristics
POSTPROCESSOR
morphological heuristics, lexeme based valency tags and semantic class potential
MORPHOLOGICAL DISAMBIGUATION
iterative context based CG disambiguation rules, based on:
word class, word form, base form, valency potential, semantic prototypes
TAGGER
"DANTAG"
SYNTACTIC MAPNING
adds lists of possible syntactic function tags / constituent markers (at word or subclauce level) to words, based on disambiguated morphology and context
SYNTACTIC DISAMBIGUATION
iterative context based CG disambiguation rules, handles:
argument structure, dependency relations at group and clause level
subclause form and function
PARSER
"DANSYN"
PROPRIUM-CG
recognition of name types using contextual CG rules
CASE ROLE-CG
context based mapping and disambiguation of semantic case roles
(Søren Harder)
PSG
generation of syntactic tree structures, using generative rewriting rules
MT-CG
context based mapping of translation equivalents Danish -> Esperanto
"DANTRAD"
spell and grammar checking
clause boundaries
teaching software
language games
corpus annotation

The DanGram system in current numbers

Lexemes in morphological base lexicon: 146.342
(equals about 1.000.000 full forms), of these:
proper names: 44839 (experimental)
polylexicals: 460 (+ names and certain number expressions)
Lexemes in the valency and semantic prototype lexicon: 95.308
Lexemes in the bilingual lexicon (Danish-English: 88.000, Danish-Esperanto: 36.000)

Danish CG-rules, in all: 6.233
morphological CG disambiguation rules: 2.678
syntactic mapping-rules: 1.701
syntactic CG disambiguation rules: 1.854
(plus 429 bilingual rules in separate MT grammars, and a smaller number of semantic case-role and proper name-rules in the semantics and name grammars)

Danish PSG-rules: 490 (for generating syntactic tree structures)
Danish Dependency-rules: ~ 267 (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 3-6 sec)
morphological analysis alone: ca. 1000 words/sec

VISL parsing tools

- Preprocessing: word- and sentence boundaries, polylexicals
- Lexicon and rule based morphological analysis: Inflexion, derivation, composita recognition
- Postprocessing: Valency and semantic potential
- Morphological contextual disambiguation (CG)
- Syntactic mapping og diambiguation (CG)
- Names CG , feature propagation CG, Case role-CG
- PSG/Dep-layer: Teaching, Arboretum, Floresta

Externally co-funded research projects

- **SHF** 1999-2001: CG, syntax & semantics (da, en, po)
- **AC/DC** 1999-?: Portuguese CG-corpora
- **Floresta** 2000-?: Portuguese treebank
- **DSL** 2001-?: Korpus90/2000 (Danish CG-corpora)
- **Arboretum** 2002-2005: Danish treebank
- **PaNoLa** 2002-2006: Integration of Nordic CG research
- **Nomen Nescio** (2003-2004), **HAREM** (2004-2005): Automatic named entity recognition
- **Nordic Treebank Network**: 2003-2005

Running CG-annotation

Da	[da]	**KS**	@SUB
den	[den]	**ART** UTR S DEF	@>N
gamle	[gammel]	**ADJ** nG S DEF NOM	@>N
sælger	[sælger]	**N** UTR S IDF NOM	@SUBJ>
kørte	[køre]	<mv> **V** IMPF AKT	**@FS-ADVL>**
hjem	[hjem]	**N** NEU P IDF NOM	@<ACC
i	[i]	**PRP**	@<ADVL
sin	[sin]	<poss> <refl> **DET** UTR S	@>N
bil	[bil]	**N** UTR S IDF NOM	@P<
så	[se]	<mv> **V** IMPF AKT	@FMV
han	[han]	**PERS** UTR 3S NOM	@<SUBJ
mange	[mange]	<quant> **DET** nG P NOM	@>N
små	[lille]	**ADJ** nG P nD NOM	@>N
dyr	[dyr]	**N** NEU P IDF NOM &ACI-SUBJ	@<ACC
på	[på]	**PRP**	@<OA
de	[den]	**ART** nG P DEF	@>N
våde	[våd]	**ADJ** nG P nD NOM	@>N
veje	[vej]	**N** UTR P IDF NOM	@P<

Cross language perspective

- VISL uses a uniform descriptive system, with consistent form and function categories, for 27 languages, handling special cases at the subcategory level
- CorpusEye offers 2 large CG-annotated multi-language corpora, allowing a certain degree of statistical standardisation (genre, lexicon etc.) across languages
  - 1. Europarl parallel corpus (da, de, en, es, fr, it, pt)
  - 2. Wikipedia corpus (da, de, en, eo, es, fr, it, pt)
- Both the annotation (e.g. np-types), search system (e.g. different statistics) and language inventory (e.g. se) can be expanded in a project-driven way

Cross SL category distribution

	da	sv	de	en	nl	GER	xx/fr	es	it	pt	ROM	fi	el
words per sentence	25.5	25.1	25.3	25.7	**23.1**	**24.9**	**27.8**	32.1	32.9	**33.2**	**32.7**	25.3	31.0
finite subclauses	3.81	3.75	3.47	3.47	3.30	3.56	**3.16**	**4.04**	3.68	3.52	3.75	**3.00**	3.72
relative clauses	1.95	2.05	1.68	1.70	1.58	**1.79**	**1.72**	**2.16**	2.10	2.07	**2.11**	**1.50**	2.09
direct object clauses	**1.11**	1.04	1.02	1.03	0.95	**1.03**	**0.85**	1.10	0.90	0.81	**0.94**	**0.78**	0.94
adverbial clauses	0.63	0.54	0.67	0.61	0.63	0.62	**0.52**	**0.70**	0.63	0.55	0.63	0.57	0.62
participial adverbial subclauses (log-5)	2.92	**2.15**	3.20	4.35	4.52	**3.43**	**3.96**	3.82	4.09	4.71	**4.21**	3.31	**4.78**
auxiliary chain parts	**3.46**	3.35	3.34	3.36	3.13	**3.33**	2.89	2.98	2.99	2.52	**2.83**	3.02	**2.77**
passive pcp2	**0.47**	0.45	0.42	0.45	0.44	0.45	0.41	**0.33**	0.34	0.39	0.35	0.44	0.39
active pcp2	1.17	1.14	1.15	**1.33**	1.07	1.17	1.12	1.22	1.20	**0.95**	1.12	1.04	1.17
infinitive	**1.43**	1.38	1.39	1.21	1.25	1.33	**0.99**	1.12	1.11	**0.93**	1.05	1.20	0.89
subjunctive/vfin	4.99	**5.58**	4.76	4.53	4.40	4.85	**4.19**	4.76	4.26	4.79	4.60	5.55	4.35
conditional	0.56	0.56	0.56	**0.62**	0.43	**0.55**	0.43	0.49	0.43	**0.40**	**0.44**	0.56	0.39
vocative	**0.04**	**0.04**	0.06	0.05	0.06	0.05	0.05	0.06	**0.07**	**0.04**	0.06	0.05	0.05
attributive	**6.70**	6.98	7.02	7.01	7.29	**7.00**	**7.26**	7.37	7.64	**8.13**	**7.71**	7.65	7.62
common nouns	**20.90**	21.26	21.00	21.33	21.35	21.2	22.07	21.37	21.09	22.14	21.5	**22.66**	21.71
finite verbs	**8.94**	8.59	8.48	8.29	8.49	**8.56**	7.57	8.18	7.78	**7.23**	**7.73**	7.83	7.86
coordinating conjunction	2.67	2.48	2.80	2.68	2.56	**2.64**	**2.74**	3.20	**3.16**	3.28	**3.21**	**2.40**	3.20
subordinating conjunct.	2.33	2.16	2.22	2.17	2.13	2.20	**1.84**	**2.35**	2.01	1.87	2.08	1.88	2.06
demonstrative	**1.96**	2.14	**2.34**	2.17	2.24	2.17	**1.99**	2.17	1.98	2.02	2.06	1.82	1.81

GER = Germanic average, ROM = Romance average, Red = high values, Blue = low values

Notables: Sentence length, inflexion vs. aux chains, subjunctive and conditional, ROM-adj vs. GER-v, ROM-coord., DK vs. ES, xx-French (shorter than even GER), politeness vocative

References

Bick, Eckhard (1997), "Internet Based Grammar Teaching", in *Datalingvistisk Forenings Årsmøde 1997 i Kolding, Proceedings,* Ellen Christoffersen & Bradley Music (red.), pp. 86-106. Kolding: 1997 Institut for Erhvervssprog og Sproglig Informatik, Handelshøjskole Syd.

Bick, Eckhard (2001). ”En Constraint Grammar Parser for Dansk”. In: Widell, Peter & Kunøe, Mette (ed.): *8. Møde om Udforskningen af Dansk Sprog*. Århus: Århus Universitet 2001.

Bick, Eckhard (2003-1), “Arboretum, a Hybrid Treebank for Danish”. In: Joakim Nivre & Erhard Hinrich (eds.), *Proceedings of TLT 2003 (2nd Workshop on Treebanks and Linguistic Theory, Växjö, November 14-15, 2003)*, pp.9-20. Växjö University Press

Bick, Eckhard (2003-2). “A CG & PSG Hybrid Approach to Automatic Corpus Annotation”. In: Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster), pp. 1-12

Bick, Eckhard (2003-3), *Grammy i Klostermølleskoven - "VISL light": Tværsproglig sætningsanalyse for begyndere.* Århus: 2002, Forlaget Mnemo

Bick, Eckhard (2004), "Grammatik for sjov: IT-baseret grammatik-læring med VISL", in *Call for the Nordic Languages: Tools and methods* (Proceedings of NorFa CALL Net Symposium Sept. 30. - Oct. 1. 2004), Peter Juel Henrichsen (red.), København: 2004

Bick, Eckhard (2005), “CorpusEye: Et brugervenligt web-interface for gramatisk opmærkede korpora”, in *10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings,* Peter Widell & Mette Kunøe (red.), pp.46-57, Århus: 2005, Århus Universitet

Christ, Oli (1994), "A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest: 1994

Dansk Sprognævn, "Kommaregler". Copenhagen: Dansk Sprognævn, pp. 17-30, København: 2004

Dienhart, John (2000), "VISL-projektet: Om anvendelse af IT i sprogundervisning og -forskning", in *At undervise med IKT,* pp. 51-70. Gylling: 2000, Narayana Press

Jandorf, Birgit Dilling (red.), *Rapport om OrdRet - en it-baseret stavekontrol,* København: 2005Maegaard, Bente et.al. (2004), “Strategisk Satsning på Dansk Sprogteknologi”, København: 2004, Statens Humanistiske Forskningsråd

Robb T. (2003) "Google as a Quick 'n Dirty Corpus Tool", *TESL-EJ* 7, 2. Available at:**http://www-writing.berkeley.edu/TESL-EJ/ej26/int.html**

Tapanainen, Pasi (1999). *Parsing in two frameworks: finite-state and functional dependency grammar*. University of Helsinki, Deparment of General Linguistics

Tapanainen, Pasi and Timo Järvinen. (1997). ”A non-projective dependency parser”. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington, D.C., April. Association for Computational Linguistics.