FrAG
A Hybrid CG Parser for French

Eckhard Bick
University of Southern Denmark
eckhard.bick@mail.dk

Outline

- Background: Research environment and data
- The FrAG parser and its modules
- Annotation scheme
- Evaluation
- Dependency vs. PSG issues

- Applications, corpus work
- Outlook
-

Background

- VISL project at University of SoutherDenmark:
  - CALL grammar for 25 languages
  - French parser project active esp. 2003-04, 2009
- CorpusEye: Corpus annotation project for ~ half of those languages, CG and treebanks
- Deep (= full tree) parsers to support this
- Open Source Constraint Grammer compiler (CG3 by GrammarSoft ApS)
- General Language Technology Perspective: MT, Grammar checking, NER

Dual Hybridity

Format hybridity: 3 parallel, but not wholly information-equivalent, output formats

(a) word based functional dependency tags (CG)

(b) VISL-style constituent trees

(c) Other treebank schemes: PENN-treebank, TIGER, dependency trees

with all formats sharing tags for syntactic function and morphological form.

Hybrid parsing/annotation process:

1.) Probabilistic Decision Tree Tagger (A. Schmid & H. Stein)

1 --> **2.)** Morphological analysis

2 --> **3.) lexicon and rule driven morphosyntactic analysis (CG)**

3 --> **4.)** shallow dependency parsing (CG)

4 --> **5.a)** function based constituent analysis (PSG)

4 --> **5.b)** full dependency (separate grammar or CG3)

FrAG Modules

Decision Tree Tagger (Schmid 1994):
probabilistic PoS tagging

Constraint grammars: rule & context based;
morphology, syntax, attachment, clause boundaries
(e.g 1.560 French rules, of these 167 correction and 270 attachment/dependency rules)

Rule compilers: vislcg3 (GrammarSoft open source)

Lexica: inflexion, valency, polylexicals, names etc.

- 65.470 lexemes:
- 6.200 verbs with valency patterns,
- 17.860 nouns with semantic prototype information, e.g. <Hprof>, <tool>)

Secondary programs: format filters, VISL's graphical tree manipulator, corpus search tools, linux editors, ...

Tokenisation

Fusion:

- polylexical prepositions, conjunctions, adverbs
  qu'est-ce_que, tout_à_fait
- Name chains
  Charles_de_Gaulle

Splitting:

- prp+art: *du, des* (disambiguated from partitive/art), *au, aux*
- Apostrophe: *n'a, c'est*

Punctuation: Used as context, sentence delimiters and parentheses as “word tokens”

Dependency, form and function

CG-level: Each text token is assigned

a function tag (subject, auxiliary, ...) and a form tag (PoS, clause type, ...)

a directed shallow CG-dependency, pointing to a head-category explicitly (@>N prenominal) or implicitly (@<SUBJ subject right of verb).

Full dependency:

number markers for full dependency (e.g. #5 = dependent of word 5)

computed from

- shallow CG-dependency
- uniqueness principle
- special secondary attachment tags (close, long, coordination)

PSG-level with constituent trees:

adds clause and group boundaries

adds explicit discontinuity and raising

creates head-function (H)

retains group-specific dependency-functions (e.g. DN for nominal groups).

30 major syntactic functions

Table 1: Syntactic functions

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

Valency potential - the lexical key to syntax

Valency lexicon: valency potential for verbs and nouns
<vt> <vdt> <ve> <på^vp> <vq> <vi-ud> <xt>, <+INF> <+på> <+num> ...

Annotation:
Valency controlled tag choices on dependents rather than structural marking
Disambiguation of valency potential markers

Example: valency-inspired Pp-nodes

- (free) **adunct adverbial** (fA): *selon lui, d'abord, il travail ici*
- (bound) **argument adverbial** e.g. **with** object relation (Ao): *mettre en place (quelque part*
- (bound) **prepositional object** (Op): *demande à qn de fair qc*

underspecified valency at group level

- **adnominal dependent** (DNmod): *les derniers points*, *la pipe du père*
- **adverbial dependent** (DAarg): *supérieur à*

Experimentally, **case roles** like Actor, Patient etc. are assigned by a special layer of CG rules, using function context, valency and lexical information handed down by the other CG-modules.

Running CG-annotation

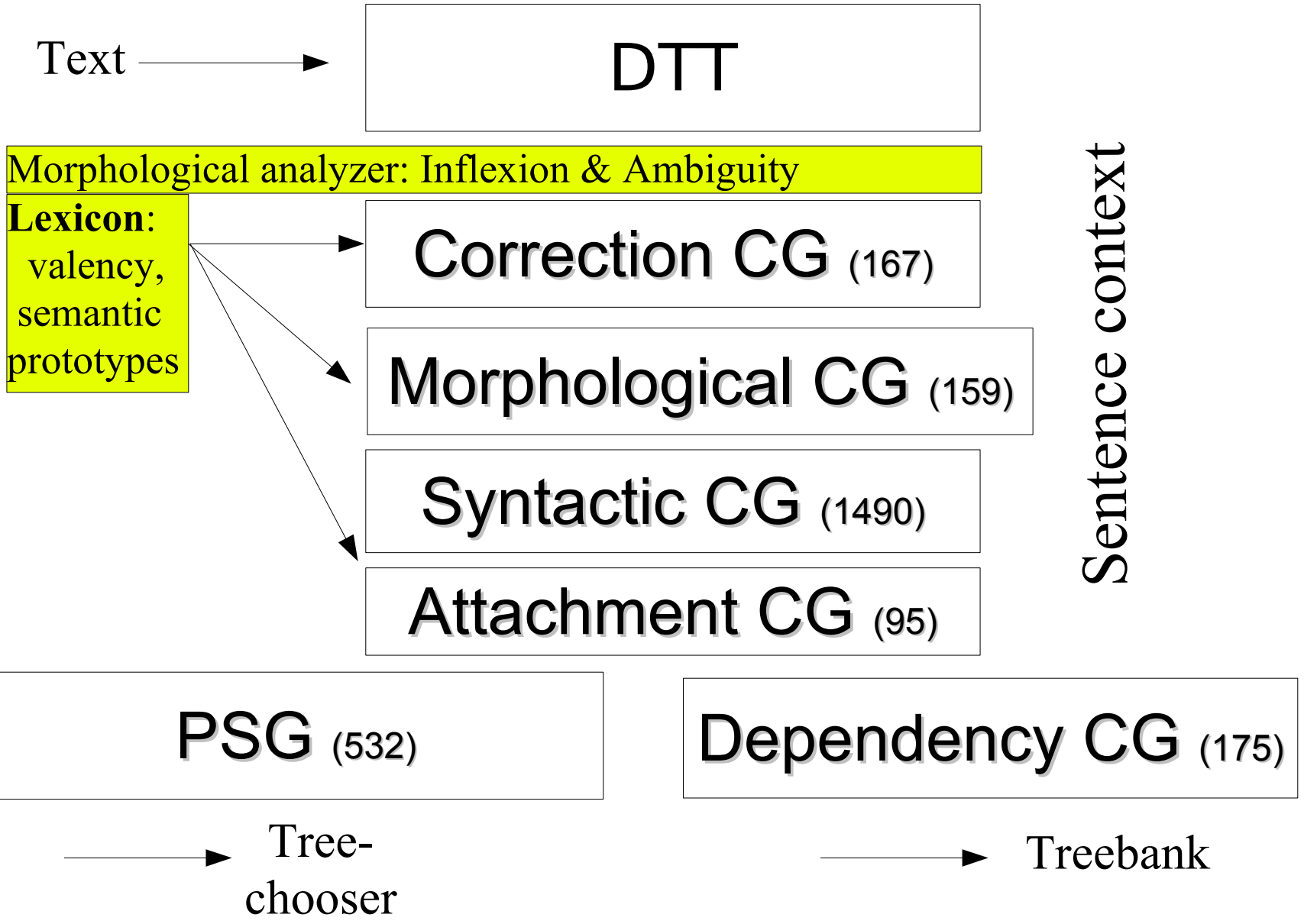
1. Il	[il]	**PERS** 3S NOM	**@F-SUBJ>**	**#1->2**
2. faudrait	[falloir]	**V** 3S COND	**@FMV**	**#2->0**
3. que	[que]	**KS**	**@SUB**	**#3->5**
4. je	[je]	**PERS** 1S NOM	**@SUBJ>**	**#4->5**
5. puisse	[pouvoir]	<aux> **V** PR 1/3S SUBJ	**@FS-<SUBJ**	**#5->2**
6. alterner	[alterner]	<mv> **V** INF	**@AUX<**	**#6->5**
7. avec	[avec]	**PRP**	**@<PIV**	**#7->6**
8. les	[le]	**ART** nG P	**@>N**	**#8->9**
9. autres	[autre]	**ADJ** nG P	**@P<**	**#9->7**

(It is necessary that I can take turns with the others.)

Une	[une] <idf>	ART	@>N	#1->2
direction	[direction]	N F S	@SUBJ>	#2->13
spéciale	[spécial]	ADJ F S	@N<	#3->2
,				#4->0
instituée	[instituer] <mv>	V PCP2 ...	@ICL-N<	#5->2
à	[à] <sam->	PRP	@<ADVL	#6->5
le	[le] <-sam> <def>	ART M S	@>N	#7->8
ministère	[ministère]	N M S	@P<	#8->6
de	[de] <np-close>	PRP	@N<	#9->8
la	[le] <def>	ART F S	@>N	#10->11
guerre	[guerre] <clb-end>	N F S	@P<	#11->9
,				#12->0
est	[être] <aux>	V PR 3S IND	@FS-STA	#13->0
chargée	[charger] <mv>	V PCP2 ...	@AUX<	#14->13
de	[de]	PRP	@<PIV	#15->14
tout	[tout] <quant>	PRON DET M S	@>N	#16->17
ce	[ce] <dem>	PRON INDP M S	@P<	#17->15
qui	[qui] <rel>	PRON INDP NOM	@SUBJ>	#18->19
concerne	[concerner] <mv>	V PR...	@FS-N<	#19->17
le	[le] <def>	ART M S	@>N	#20->21
personnel	[personnel]	N M S	@<ACC	#21->19

(A special administration, created by the Ministry of War, has been charged with everything that concerns the personel.)

How to get from text to tree?



Filtered DTT-output (probabilistic)

Enter French text to parse:

Je crois, que j'ai eu de la chance.

Parser: Probabilistic Tagger Visualization:

Je [je] **PRON** <pers> <conj>
crois [croire] V PR IND
,
que [que] **CONJ** KS
j' [je] **PRON** <pers> <conj>
ai [avoir] V PR IND <aux>
eu [avoir] V PCP2
de [de] **PRP**
la [le] **ART** <def>
chance [chance] N

Constraint Grammar output

Enter French text to parse:

Je crois que j'ai eu de la chance.

Parser: CG-Parser Visualization: Default

Je [je] **PRON** PERS 1S NOM @SUBJ>
crois [croire] <mv> **V** PR 1/2S IND @FMV
que [que] KS @SUB
j' [je] **PRON** PERS @SUBJ>
ai [avoir] <aux> **V** PR 1S IND @FS-<ACC
eu [avoir] <mv> **V** PCP2 M S AKT @ICL-AUX<
de=la [de+le] <idf> **ART** F S @>N
chance [chance] N F S @<ACC

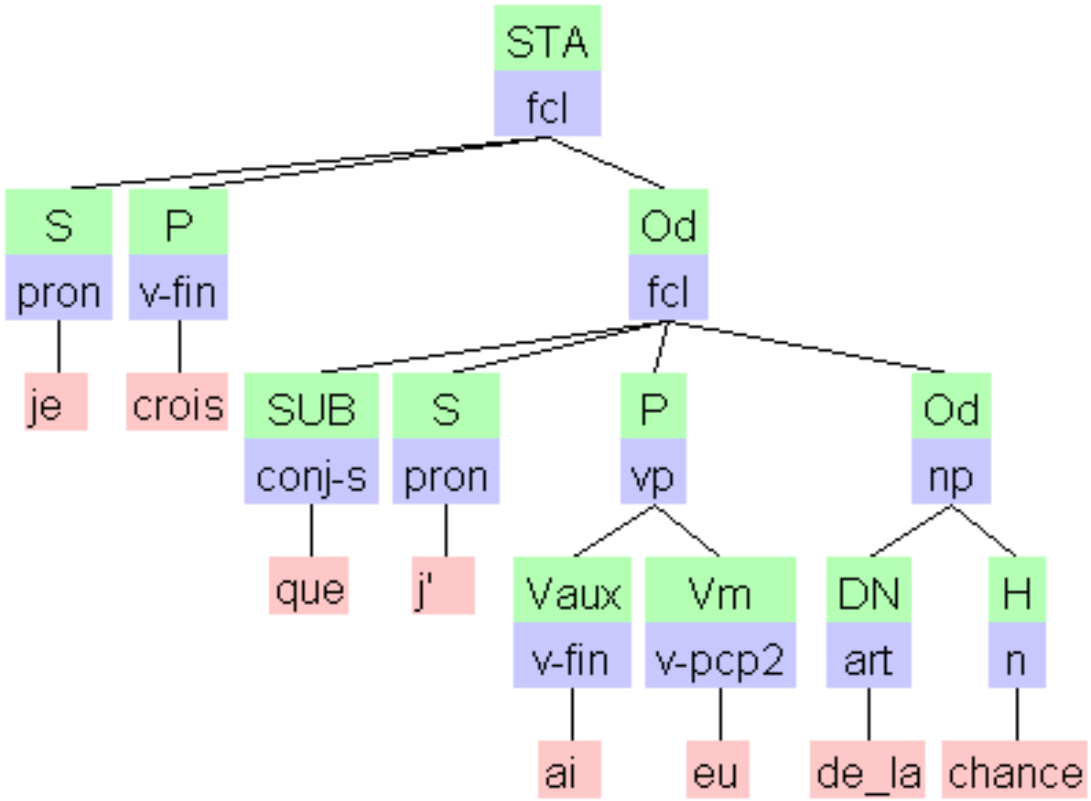
Constituent trees (PSG-output)

```
.....
Od:fc1
=S:np
==DN:art('le' <def> F S)       La
==H:n('télévision' F S) télévision
==DN:fc1
===Od:pron-rel('que' <rel> INDP ACC)    qui
===S:pron-pers('nous' PERS 1P nC)      nous
===P:vp
====Vaux:v-fin('avoir' PR 1P IND)      avons
====Vm:v-pcp2('proposer' F S AKT)      proposée
===fA:pp
====H:prp('à' <sam->)   à_
====DP:np
=====DN:art('le' <-sam>  M S)   _le
=====H:prop('CSA' M S)  CSA
=P:vp
==Vaux:v-fin('être' FUT 3S IND) sera
==Vm:v-pcp2('mettre' F S PAS)   mise
.....
```

FUNCTION:form
EDGES:nodes/terminals

indentation for depth

Constituent trees (graphical)



Evaluation 1

CG-annotation for French Europarl data

(1.790 words)

	Recall	Precision	F-score
Word classes (CG)	98.7 %	98.7 %	98.7
Syntactic functions	93.7 %	92.5 %	93.1

Comparison: DTT-stage alone: 97.5% F-score for PoS

Coparison: 2003 version on news text: 17.500 words, long sentences (28 words av.)
F-Score 97, DTT alone 95.7

mature Constraint Grammars: > 95% syntactic accuracy, ca. 99% PoS accuracy

- French FSP (Chanod & Tapanainen 1997), Portuguese/Danish CG (Bick 2003)

---

[1] separately counting tenses, participles and infinitive

[2] including subclause functions, but without making a distinction between free and valency bound adverbials

Evaluation 2

CG-annotation for Wikipedia

(1.714 words, 1911 tokens)

	Recall	Precision	F-score
Edge label/functions	96.20%	96.20%	96.2
Dependency links	95.90%	95.90%	95.9

Comparison: Probabilistic ML parsers

- Crabbé et al. (2009): edge label F-score 87.2 (66.4 external EASY)
- Schulter & van Genabith (2008): LFG-derived SVM-system F=86.73
- Arun & Keller (2005): unlabelled dependency F-score 84.2
- Candito et al. (2009): unlabelled dependency F-score 90.99

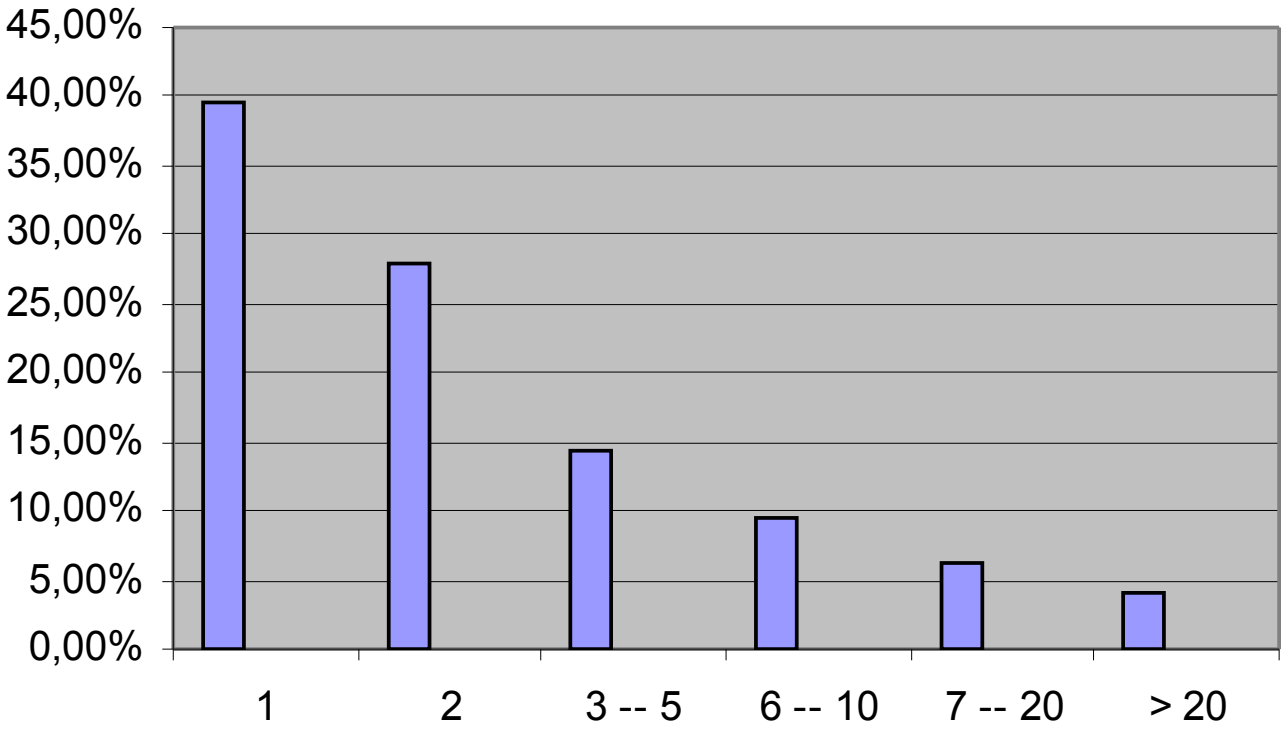
---

[1] separately counting tenses, participles and infinitive
[2] including subclause functions, but without making a distinction between free and valency bound adverbials

full tree-creation: PSG first vs. Dependency first

- PSG is less robust than dependency:
  PoS/funciton errors affect whole trees
  ungrammatical sentences are worse in PSG
- Coordination and ellipsis is descriptively more natural in constituent grammar, but methodologically easier in dependency grammar
  - e.g. missing subject or coordinator
- Discontinuous constituents are harder to handle in PSG than CG:
  - verb chain "brackets"
  - topic/focus fronting
  - raising
- PSG has time-space problems for complex input

Percentage of forests with n trees



CG-to-Tree conversion: Solutions

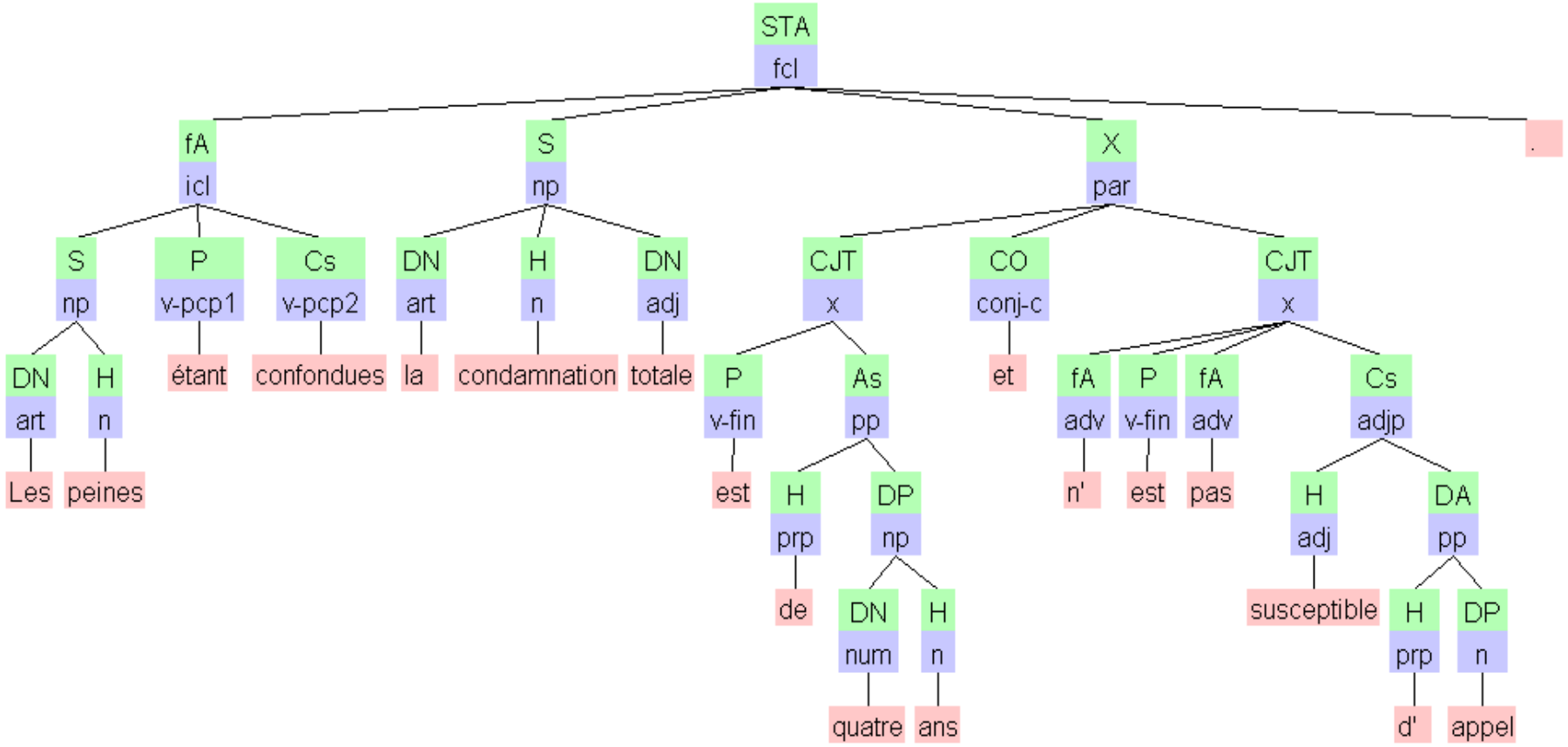
First step: Attachment CG:
- attachment markers: <np-close>, <np-long>
- Coordination specifiers: <co-subj>, <co-postnom> ....

Methodological ordering
- Dependency before constituent trees

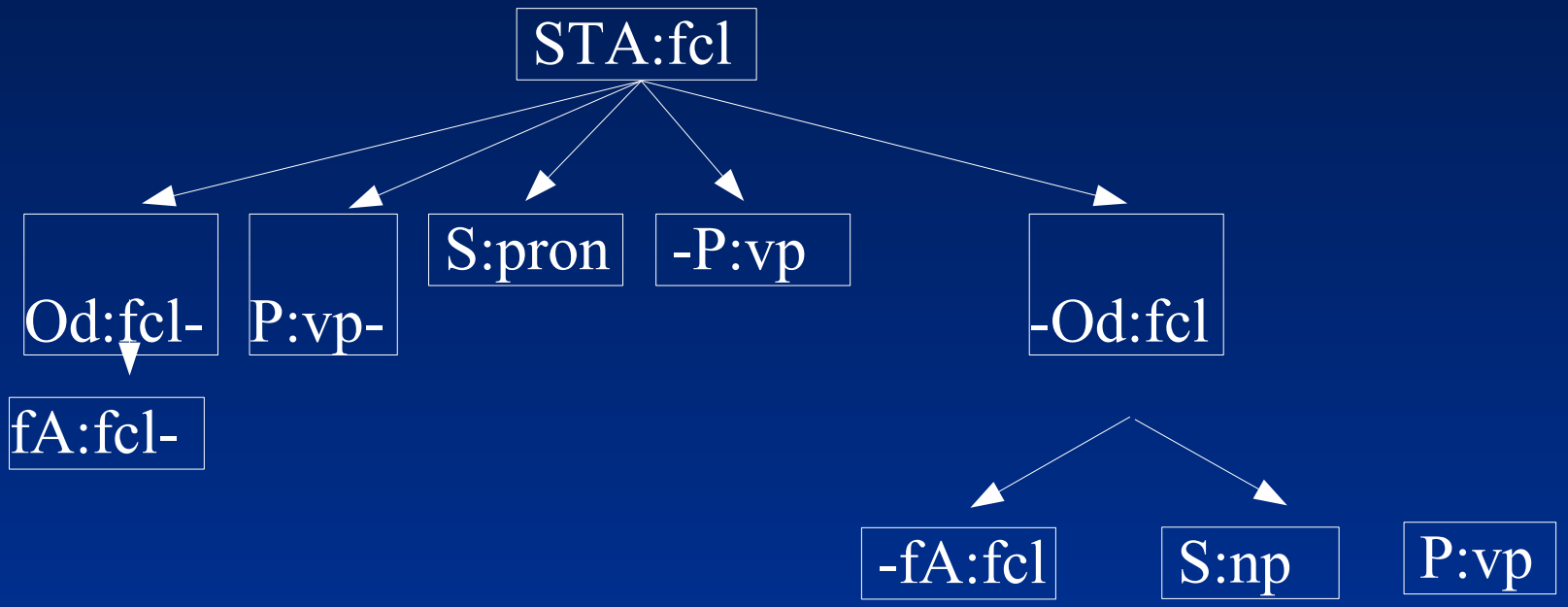
Rule ordering
- ordered attachment inside out
- coordination last

Descriptive solutions
- Stacking of "undefined" coordinated constituents
- Discontinuity marked with constituent halves

Stacking:
“Dummy” nexus conjuncts

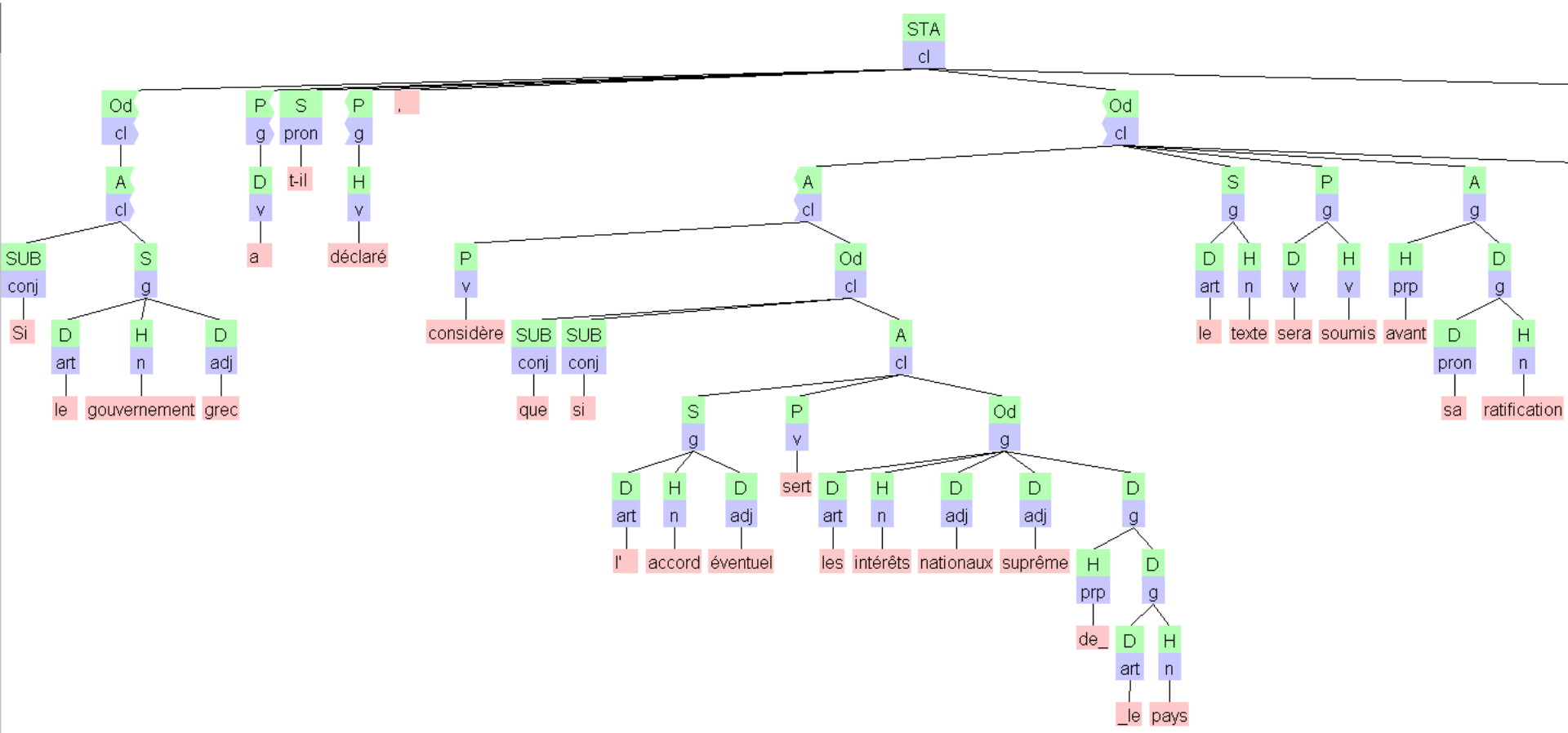


Discontinuity

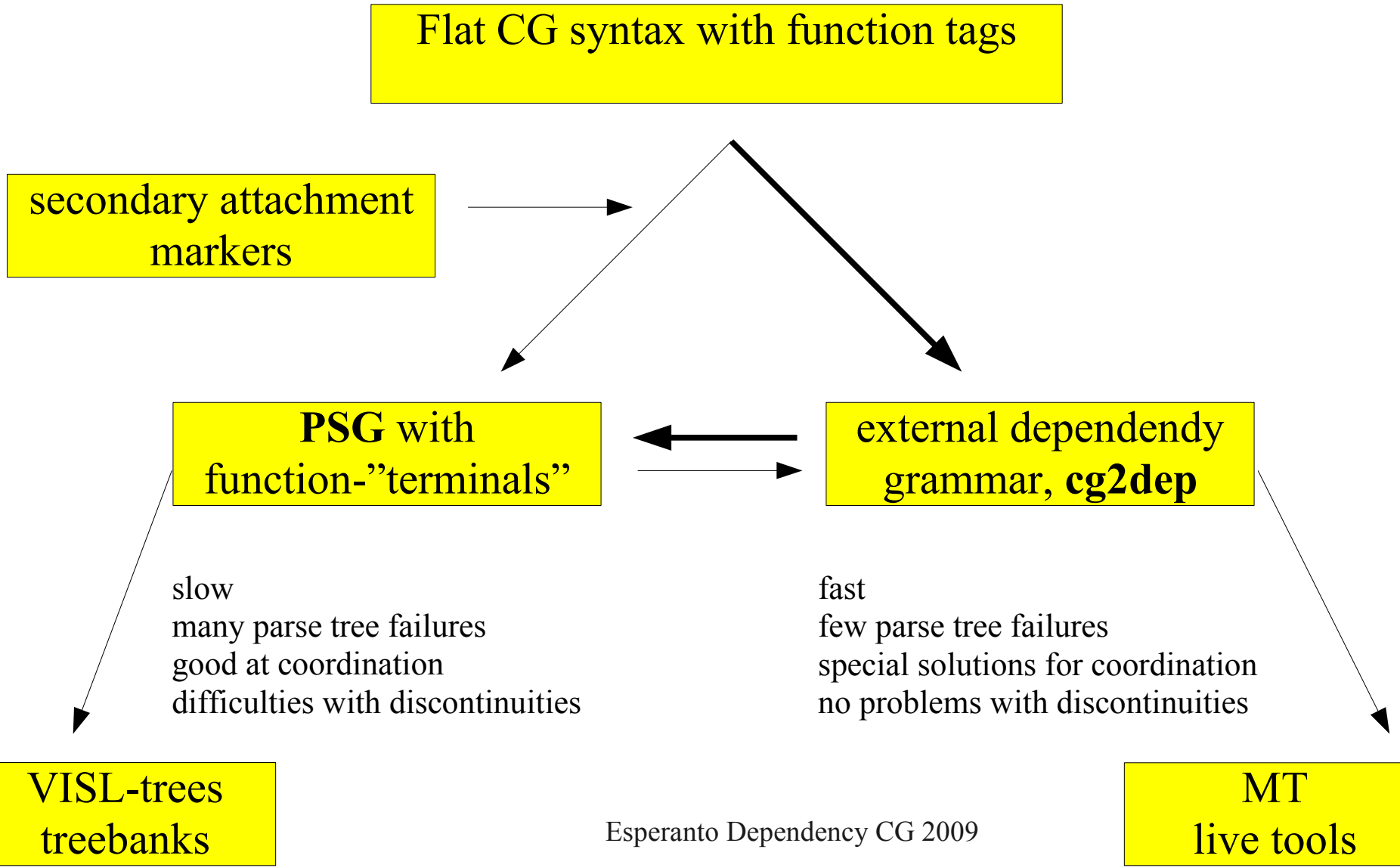


Si le gouv...	*a*	*t-il*	*declaré*	*considère que ...,*	*le texte*	*sera soumis*
FS-ADVL>>	*FMV*	*SUBJ*	*AUX<*	*FMV*	*FS-<ACC*	*SUBJ>*	*FAUX AUX<*

Discontinuity



Tree structures



Creating Dependencies in CG-3

```
SETPARENT (@>N) (0 (ART DET))  TO (*1 (N)) ;
SETPARENT (@P<) TO (*-1 (PRP)) ;
    = SETCHILD (PRP) TO (*1 @P<) ;
SETPARENT (@FS-N<) TO (*-1 @SUBJ> LINK *-1 N)  ;
```

- create dependencies on the fly
- change existing dependencies
- circularity
  - a rule won't be applied if it introduces circularity
  - however, if there IS circularity further up in the ancestor chain from a previous module, then it will be accepted

Using Dependencies

SELECT (%hum) (0 @SUBJ) (p <Vcog>)
    -> assign +HUM to subjects of cognitive verbs

SELECT (@ACC) (NOT s @ACC)
    -> uniqueness principle

(*-1 N LINK c DEF)
    -> definite np recognized through dependent

ADD (§AG) TARGET @SUBJ
    (p V-HUM LINK c @ACC LINK 0 N-NON-HUM) ;

Dependency relations

- in a rule, dep-relations (letters) replace positions (numbers)
  - Parent/Mother (p)
  - Child/Daughter (c)
  - Sibling/Sister (s)
- Complex relations can be expressed as combinations:
  - Niece: s LINK c (c LINK s = 2 c-tests)
  - Aunt: p LINK s (s LINK p = p)
  - Cousin: p LINK s LINK c

Operators

NOT regards relation existence, not tags

- NOT c @>N = no prenominal
- c @>N LINK NOT 0 P = at least one prenominal child that isn't plural (e.g. grammar checking for agreement)

* means **deep scan** of all ancestors (*p) or offspring (*c)

C means *all-relations-match,* not *all-readings-match*

- sC (P) = all siblings have a plural reading (but possibly others)
- s (P) LINK 0C (P) = there is a sibling with only plural readings

S means *and-self*

- *pS (@FS-N<) = if self or any ancestor is marked relative clause (good for verb chain testing where you don't know if you are looking at the first or later elements)

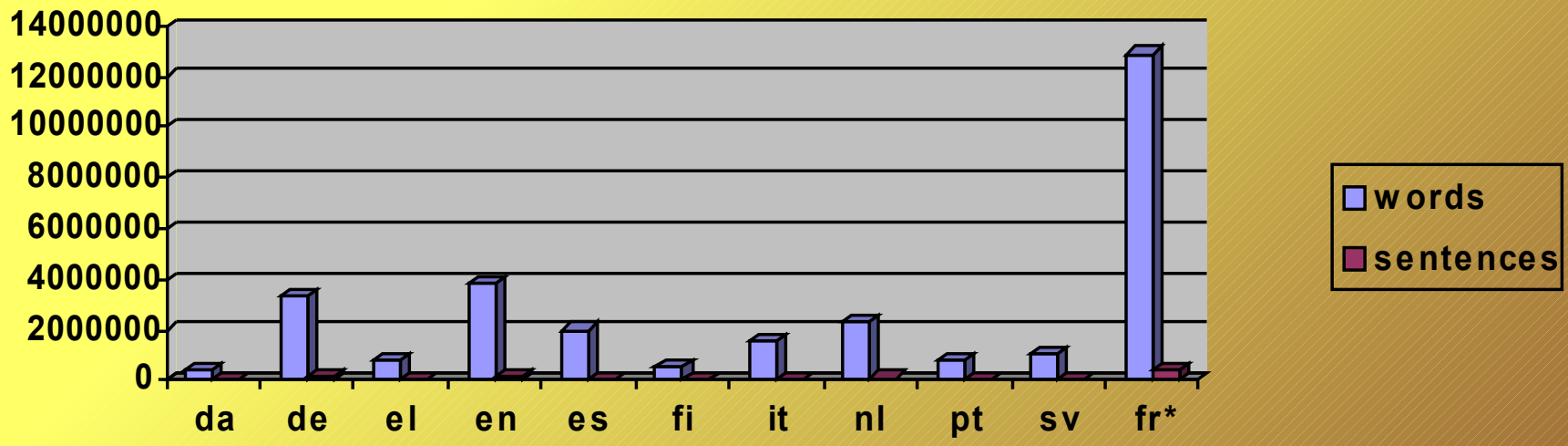
Applications

- internet-based grammar-teaching (VISL)
  - cross-language annotation scheme
  - compatible with treebanks in 25 languages
- syntactic corpus annotation
  - ANANAS-corpus (Salmon-Alt 2002)
  - l'Arboratoire treebank (manually revised)
  - Europarl annotation (28 mill. words)

The French Europarl Corpus

* 29 million words of Parliamentary debates, original or translated
* One of 11 similar corpora in the different European languages
* Freely available at http://www.isi.edu/~koehn/europarl/

Distribution of different SL in the French part of Europarl:



Cross SL category distribution (the pivot of the talk)

	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

Statistical musings

Does it make sense to statistically evaluate a corpus that has been automatically annotated, but not manually revised?

Yes, for a PoS category with a frequency of 10% and a tagger with an error rate of 1.3%, the error margin is probably(only) 9.87%-10.13%, even with all errors stemming from this category, it would only vary between 8.7% and 11.3%.

Does it make sense to compare languages through a (French) translation filter?

Yes and no – It may seem innovative, but on the other hand, French functions as a kind of neutralizing filter for arbitrary descriptive differences (traditionally varying category definitions, for instance)

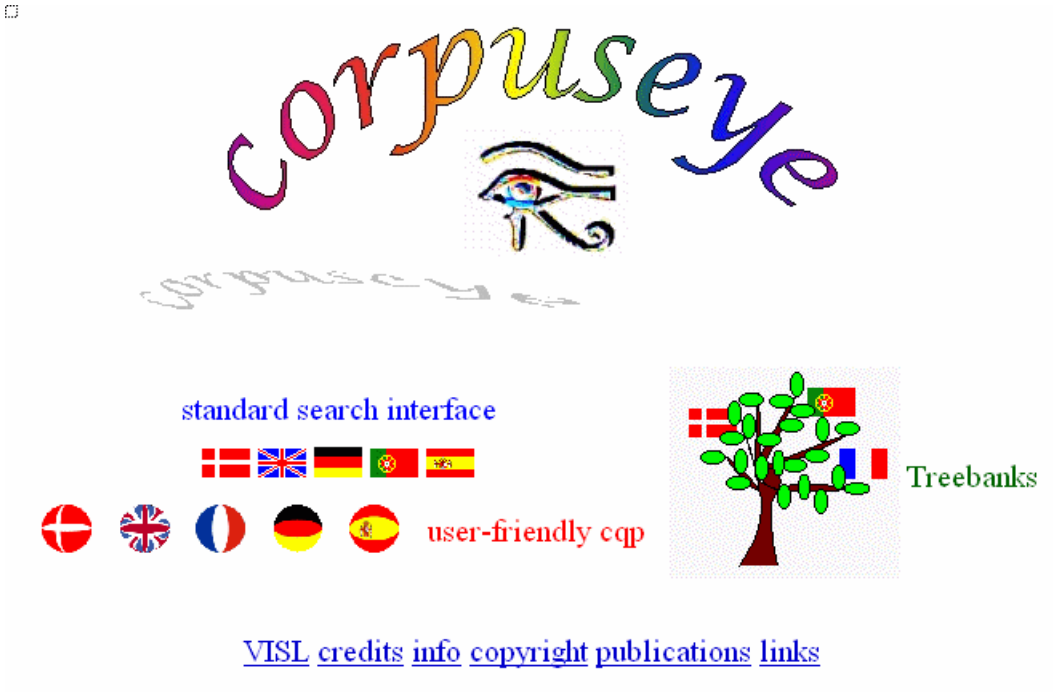
Are all SL speakers native speakers?

No. A portion of the French sources in the corpus may actually be second language speakers preferring their own French to a translated version. The same may be true for many English sources.

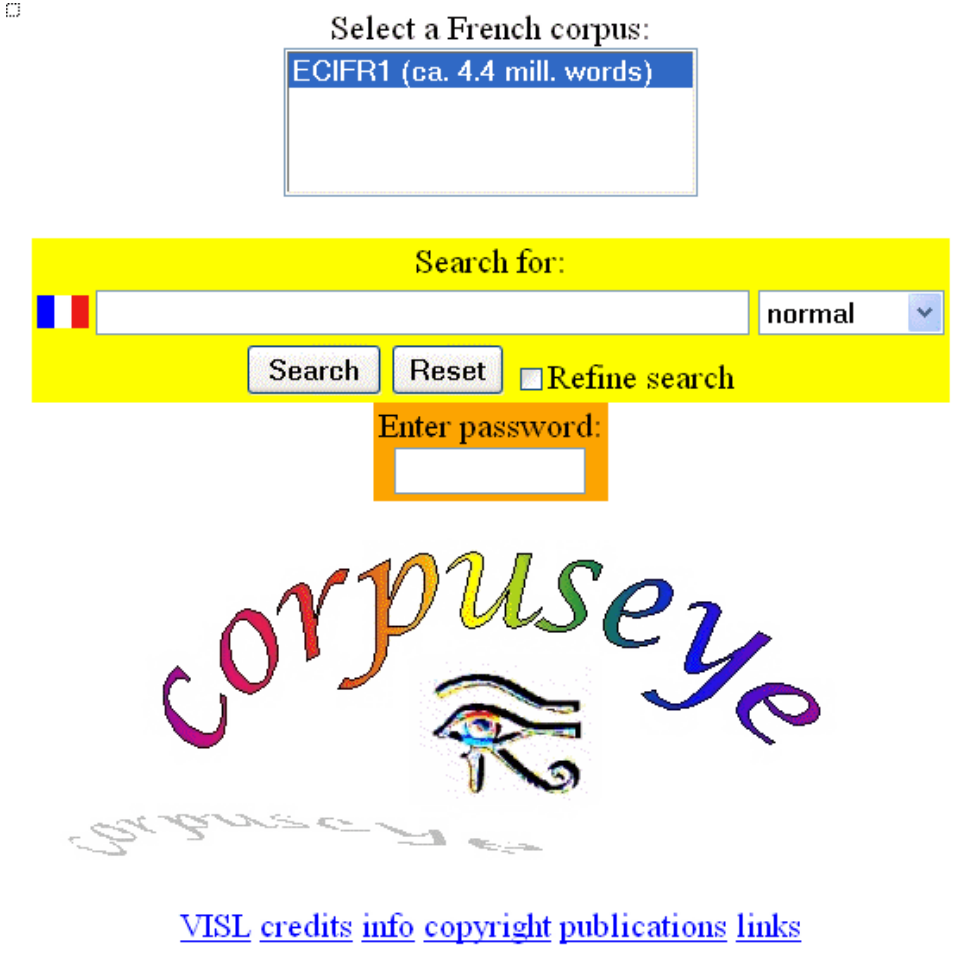
supported formats

- native VISL (indented vertical trees)
- TIGER format, both constituent and depency
- MALT dependency format (Nivre)
- PENN treebank
- XML many external versions by different research groups
- MySQL databases

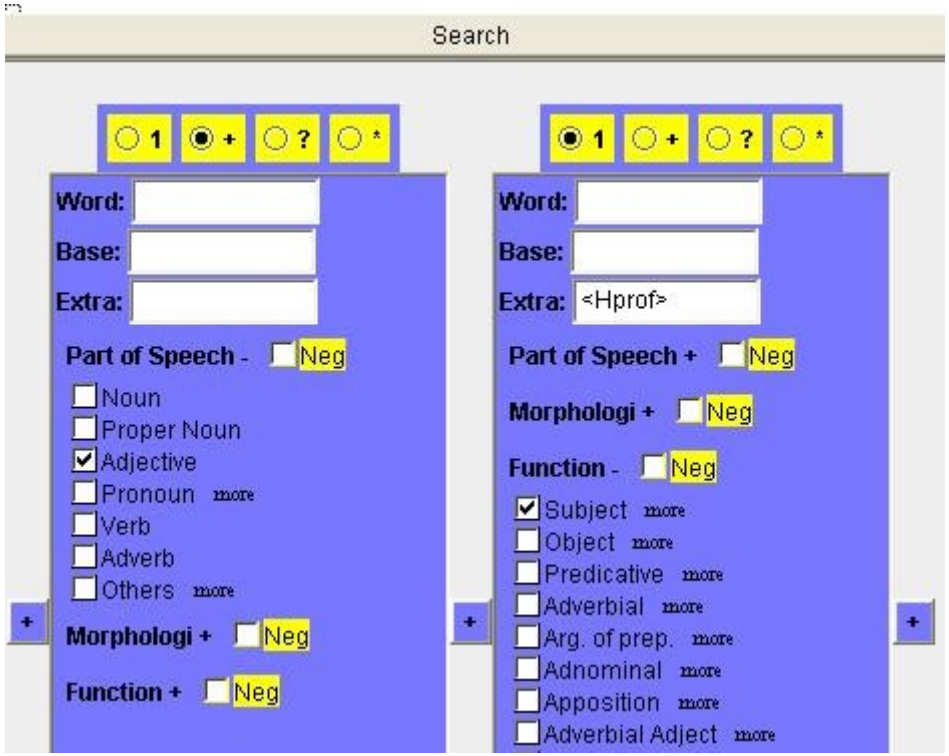
Corpus interface - top
http://corp.hum.sdu.dk



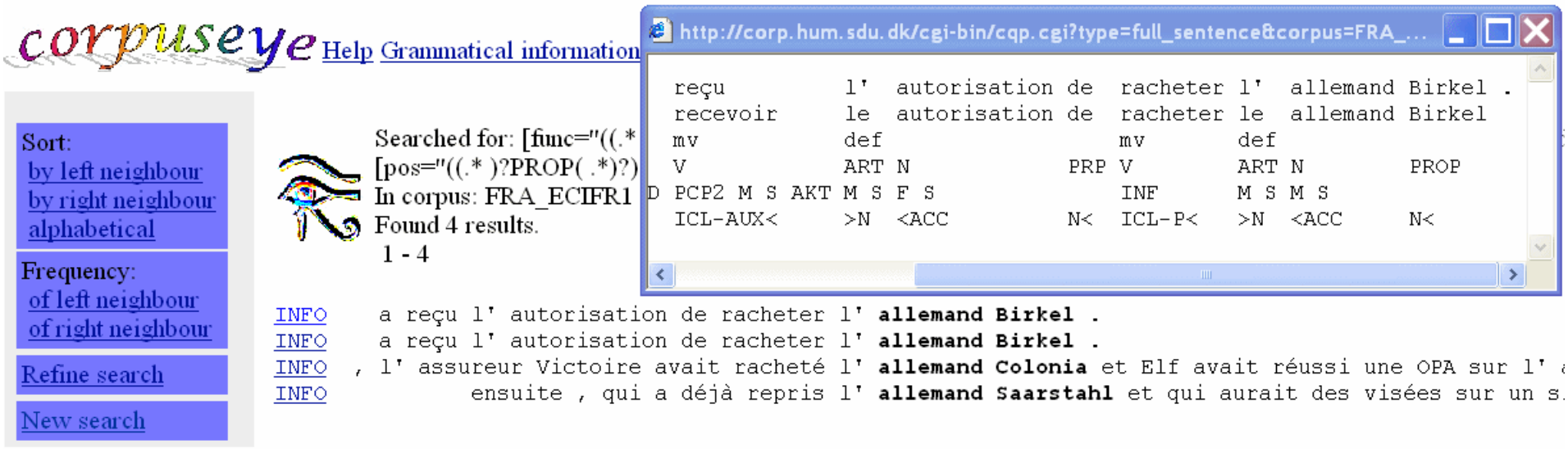
Corpus interface: French CG



Cqp-search with menus

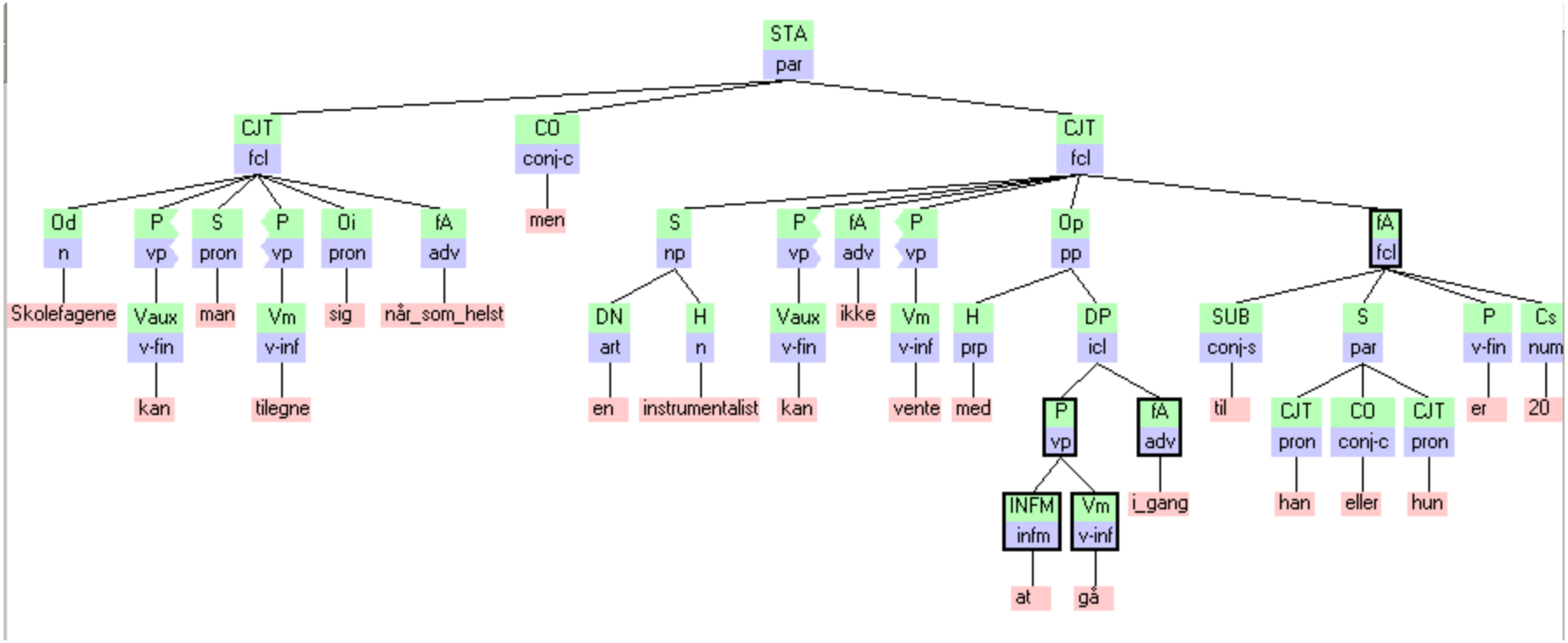


Cqp-search: Results (“allemand” @OBJECT + PROP)



Treebank search results as syntactic tree structures

P:vp~_fA:adv_fA:fcl



Outlook

- experiments with different hybridisation schemes
- integrate a from-scratch PoS CG with DTT choices to guide heuristic CG-rules
- (human) arbitration or specialist correction rules based on systematic differences between output from the 2 systems
- rule weighting based on FraG-annotated corpora

VISL

eckhard.bick@mail.dk

parsers: beta.visl.sdu.dk

corpus search: corp.hum.sdu.dk

teaching: visl.sdu.dk

Discontinuity

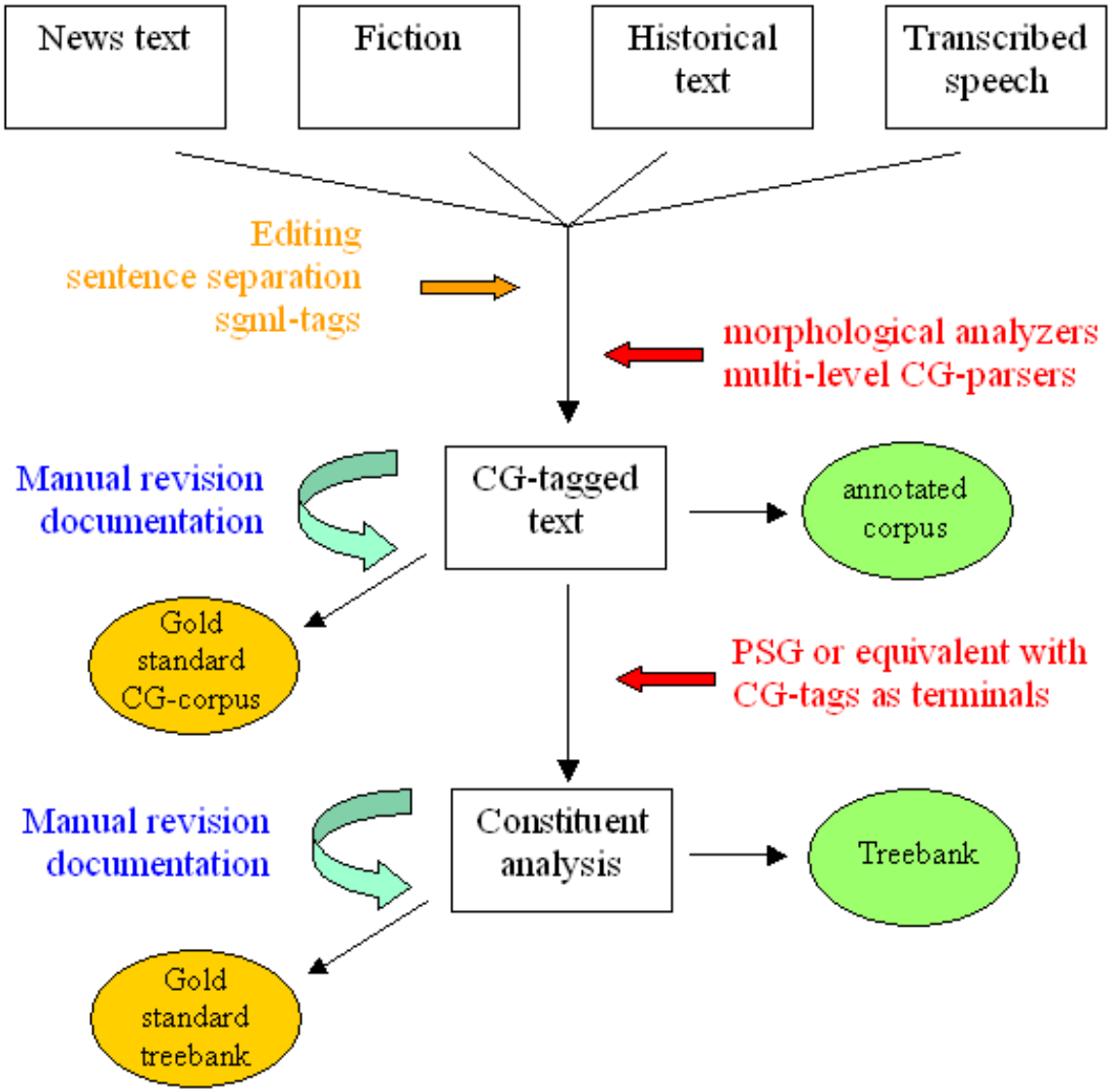
Discontinuity (crossing branches) is not a very rare feature in French, but theory dependent to a certain degree

- expressed by double-arrow-dependents in CG (meaning "cross over the next legal head, to another word of compatible type)
- expressed by "broken node"-markers in the constituent trees.
- Can be a matter of descriptional tradition:

  ne ... pas (discontinuous adv?)

  a-t-il vendu .... (auxiliaries as clause or vp constituents?)

Corpus annotation



Evaluation 2:
Constituent trees for French (PSG)

For French (FrAG)

- 45% well-formed trees on raw text
- Speed: 3000 words/sec (all CG-levels)
  37 words/sec (CG-to-tree)

for Danish (DanGram)

- 50-75% well-formed trees on raw text
- 95% well-formed trees on corrected CG-input
- 0.8% attachment errors on corrected CG-input

Choosing trees

For every sentence, a heuristic *tree chooser* program creates a priority list for surviving ambiguous trees, drawing on a variety of complexity measures:

- embedding depth
- coordination flatness
- discontinuity.

Two treebank building strategies for investing time at this point:

- Proof-reading the chosen tree
- Trust there will be one correct tree in each forest (at least with corrected CG-input and a good language-specific PSG), and therefore inspect the whole forest

Experiment: 6.800 sentences in raw, unrevised cg-format were psg-processed

- 3.191 sentences received well-formed (complete) analyses
- 3.709 sentences resulted in "fragmented" (partial) trees.

Of wellformed trees (Danish in parenthesis):

- 67% (40%) - 1 analysis, 17% (28%) - 2 analyses, 4.2% > 20 analyses
- largest forest: 192 (864) trees.

FrAG annotation scheme

Though most syntactically annotated corpora are intended as reference data for broad syntactic research in a given language, it is difficult to please all users, and a methodological or descriptional bias towards one linguistic theory or other is all but unavoidable.

"Classical" treebanks: e.g. Penn and SUSANNE treebanks

based on bracketing structure, but enriched with function labels

Dependency Grammar: Czech PDT, dep. treebanks for Turkish , Russian, Danish and Italian

Descriptional interdependency between NLP-tools and treebanks:

HPSG: Dutch Alpino-Treebank, Bulgarian BulTreeBank

LFG: Spanish UAM Treebank, PSG/Dependency TIGER-treebank for German

FrAG output comes in two parallel flavours

(a) a dependency parse with word based CG-annotation

(b) a PSG-treebank with constituent annotation (following VISL conventions)

Both versions allow crossing branches/discontinuity and specify function as well as structure and both can be converted into graphical formats

Inventory of grammatical categories follows the cross-language VISL standardisation scheme . Every node receives both a form and a function label, e.g. S:np (subject noun phrase) or fA:pp (free adverbial prepositional phrase)

Format filtering: TIGER (used as a Nordic standard), PENN (used for t-grep2)