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
• 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 Grammar 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

<table>
<thead>
<tr>
<th>@SUBJ</th>
<th>subject</th>
<th>@CO</th>
<th>coordinator</th>
</tr>
</thead>
<tbody>
<tr>
<td>@ACC</td>
<td>direct (accusative) object</td>
<td>@SUB</td>
<td>subordinator</td>
</tr>
<tr>
<td>@DAT</td>
<td>indirect (dative) object</td>
<td>@APP</td>
<td>apposition</td>
</tr>
<tr>
<td>@PIV</td>
<td>prepositional object</td>
<td>@N&lt;PRED</td>
<td>predicking postnominal</td>
</tr>
<tr>
<td>@SC</td>
<td>subject complement</td>
<td>@N&lt;</td>
<td>postnominal dependent</td>
</tr>
<tr>
<td>@OC</td>
<td>object complement</td>
<td>@N&lt;PRED</td>
<td>predicking postnominal</td>
</tr>
<tr>
<td>@SA</td>
<td>subject related argument adverbial</td>
<td>@&gt;A</td>
<td>adverbial pre-dependent</td>
</tr>
<tr>
<td>@OA</td>
<td>object related argument adverbial</td>
<td>@A&lt;</td>
<td>adverbial post-dependent</td>
</tr>
<tr>
<td>@MV</td>
<td>main verb</td>
<td>@P&lt;</td>
<td>argument of preposition</td>
</tr>
<tr>
<td>@AUX</td>
<td>auxiliary</td>
<td>@&gt;&gt;&gt;P</td>
<td>raised/fronted @P&lt;</td>
</tr>
<tr>
<td>@ADVL</td>
<td>adverbial adjunct</td>
<td>@INFM</td>
<td>infinitive marker</td>
</tr>
<tr>
<td>@AUX&lt;</td>
<td>argument of auxiliary</td>
<td>@VOK</td>
<td>vocative</td>
</tr>
<tr>
<td>@PRED</td>
<td>predicative adjunct</td>
<td>@FOC</td>
<td>focus marker</td>
</tr>
</tbody>
</table>
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 faire 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.
(It is necessary that I can take turns with the others.)
Une direction spéciale instituée à la ministère de la guerre, est chargée de tout ce qui concerne le personnel.

(A special administration, created by the Ministry of War, has been charged with everything that concerns the personnel.)
How to get from text to tree?

Text $\rightarrow$ DTT

Morphological analyzer: Inflexion & Ambiguity

Lexicon: valency, semantic prototypes

Correction CG (167)

Morphological CG (159)

Syntactic CG (1490)

Attachment CG (95)

PSG (532)

Dependency CG (175)

Tree-chooser $\rightarrow$ Treebank
Filtered DTT-output (probabilistic)

Enter French text to parse:
Je crois, que j'ai eu de la chance.

Parser: Probabilistic Tagger

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
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<
delà [de+le] <idf> ART F S @>N
chance [chance] N F S @<ACC
Constituent trees (PSG-output)

........
Od:fcl
=S:np
  —DN:art('le' <def> F S)    La
  —H:n('télévision' F S) télévision
  —DN:fcl
  ——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
........
Constituent trees (graphical)
Evaluation 1

CG-annotation for French Europarl data

(1.790 words)

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word classes (CG)</td>
<td>98.7 %</td>
<td>98.7 %</td>
<td>98.7</td>
</tr>
<tr>
<td>Syntactic functions</td>
<td>93.7 %</td>
<td>92.5 %</td>
<td>93.1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge label/functions</td>
<td>96.20%</td>
<td>96.20%</td>
<td>96.2</td>
</tr>
<tr>
<td>Dependency links</td>
<td>95.90%</td>
<td>95.90%</td>
<td>95.9</td>
</tr>
</tbody>
</table>

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:
PSPG first vs. Dependency first

- PSG is less robust than dependency: PoS/function 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.
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
Si le gouv... a t-il déclaré considère que ..., le texte sera soumis
Discontinuity
Tree structures

Flat CG syntax with function tags

Secondary attachment markers

PSG with function-”terminals”

external dependency grammar, cg2dep

VISL-trees
treebanks

MT
live tools

slow
many parse tree failures
good at coordination
difficulties with discontinuities

fast
few parse tree failures
special solutions for coordination
no problems with discontinuities

Esperanto Dependency CG 2009
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)

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

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
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 dependency
- MALT dependency format (Nivre)
- PENN treebank
- XML many external versions by different research groups
- MySQL databases
Cqp-search with menus
Sought: [func="((.*)@OBJECT+PROP(.*)?)"

In corpus: FRA_ECIF1

Found 4 results.

1 - 4

http://corp.hum.sdu.dk/cgi-bin/cqp.cgi?type=full_sentence&corpus=FRA_...

reçu l' autorisation de racheter l' allemand Birkel.

recoyer le autorisation de racheter le allemand Birkel

mv def mv def

V ART N PRE V ART N FROP

PCP2 M S ART M S F S INF M S M S

ICL-AUX< N <ACC N< ICL-P< N <ACC N<

a reçu l' autorisation de racheter l' allemand Birkel.

a reçu l' autorisation de racheter l' allemand Birkel.

l' assureur Victoire avait racheté l' allemand Colonia et Elf avait réussi une OPA sur l'

ensuite, qui a déjà repris l' allemand Saarstahl et qui surait des visées sur un s
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
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 desciptional tradition:
  - ne ... pas (discontinuous adv?)
  - a-t-il vendu .... (auxiliaries as clause or vp constituents?)
Corpus annotation

- News text
- Fiction
- Historical text
- Transcribed speech

Editing
- sentence separation
- sgml-tags

CG-tagged text

- morphological analyzers
- multi-level CG-parsers

Manual revision documentation

- Gold standard CG-corpus

Constituent analysis

PSG or equivalent with CG-tags as terminals

- annotated corpus
- Treebank

Manual revision documentation

- Gold standard treebank
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

Descriptive 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)