EASY, Evaluation of Parsers of French:what are the results?

P. Paroubek*, I. Robba*, A. Vilnat*, C. Ayache**

LREC2008 Marrakech

*

**

General presentation

EASY: Sytactic Parser Evaluation

- 1 of the 8 evaluation campaigns of the EVALDA platform, which itself is part of the TECHNOLANGUE program
- 5 corpus providers, 12 participants, 15 runs

The steps

1. at first:
   - to define the annotation
   - to collect and to annotate the corpora
   - to modify the parsers to fulfill the demands of EASY
2. to define the evaluation measures
3. to evaluate the parser results
4. to combine the results of the parsers

Outline

1. Corpus
2. Annotation of the reference
3. Evaluation measures
4. Performance
5. First ROVER test
6. Conclusion and Perspective

Corpus

Different linguistic types

- newspaper articles from *Le Monde* (as usual...)
- literary texts from ATILF databases
- medical texts, for specialized texts
- questions, with EQUER, a specific syntactic form
- manually transcribed parliamentary debates,
- "controlled" web pages and e-mails, to go further in direction of hybrid forms
- oral transcriptions

Globally :

- 40,000 sentences
- 770,000 words

Annotation of the reference

Choice made with all the participants

- small, not embedded constituents
- dependencies relations

6 kinds of constituents

- GN for Noun Phrase, as *le petit chat*,
- GP for Prepositional Phrase, as *de la maison* or *comme eux*,
- NV for Verb Kernel, including clitics as *j'ai*, or *souffert*,
- PV for Verb Kernel introduced by a Preposition, as *de venir*,
- GA for Adjectival Phrase, used for postponed adjectives in French, which are not included in GN,
- GR for Adverb Phrase as *longtemps*

Annotation of the reference : the relations

14 kinds of dependencies

- SUJ_V (subject),
- AUX_V (auxiliary),
- COD_V (direct object), CPL_V (verb complement) and MOD_V (verb modifier) for the different verb complements,
- COMP (complementor),
- ATB_SO (attribute of the subject or of the object),
- MOD_N, MOD_A, MOD_R, MOD_P (modifier respectively of the noun, the adjective, the adverb or the proposition),
- COORD (coordination),
- APP (apposition),
- JUXT (juxtaposition).

Annotation of the reference: an example from literary corpus

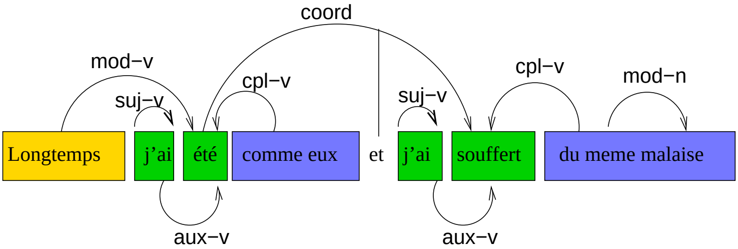

Figure: Tentative translation:*For a long time, I have lived as they do, and I suffered from the same unease*

Evaluation measures

Precision, recall and f-measure

- for constituents
- for relations
- for both of them

For each parser

- for each kind of constituent
- for each relation
- for each genre of sub-corpus
- or globally

Evaluation measures: which comparisons?

Different equality measures between two text spans from R (reference) and H (hypothesis)

- EQUALITY: $H = R$, the less permissive
- UNITARY FUZZINESS $|H \backslash R| \leq 1$
- INCLUSION: $H \subset R$
- BARYCENTER: $\frac{2 * |R \cap H|}{|R| + |H|} > 0.25$
- INTERSECTION: $R \cap H \neq \emptyset$, the most lenient

Evaluation measures: which comparisons?

Two constituents are considered equal if

- they have the same type,
- they have equal text spans.

Two relations are considered equal if

- they have the same type,
- their respective source and target have equal text spans.

Evaluation measures for constituents: global results

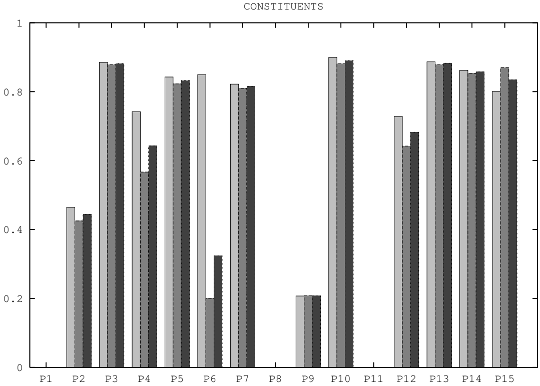

Figure: Results of the 15 parsers for constituents in precision/recall/f-measure (in this order), globally for all sub-corpora and all annotations together.

Evaluation measures for relations: global results

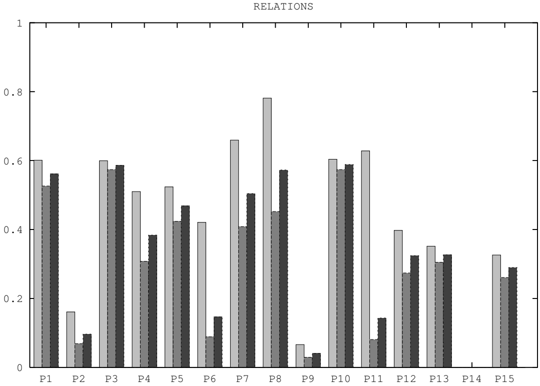

Figure: Results of the 15 parsers for relations in precision/recall/f-measure (in this order), globally for all sub-corpora and all annotations together.

Parser obtaining the best precision

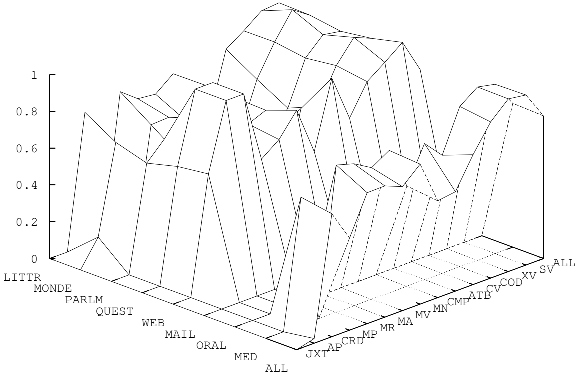

Figure: Results for relations of the parser obtaining the best **precision** measure

Parser obtaining the best recall

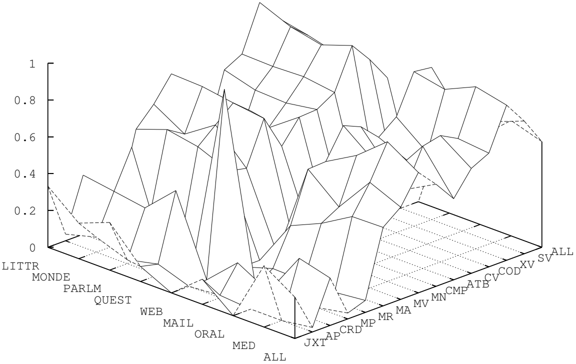

Figure: Results for relations of the parser obtaining the best **recall** measure

Parser obtaining the best f-measure

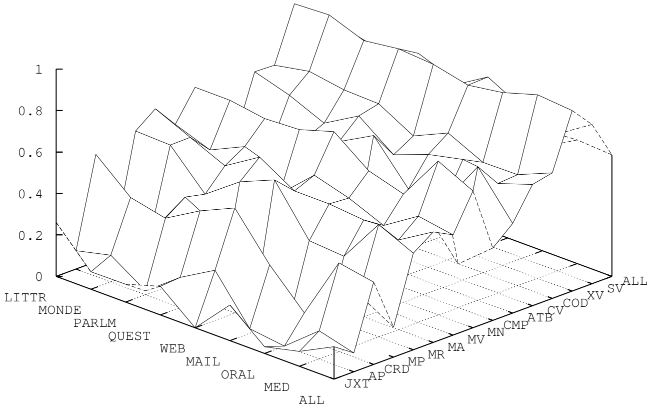

Figure: Results for relations of the parser obtaining the best **f-measure**

First conclusions

First results interesting:

- relations: best systems average f-measure near 0.60,
- high variability of results for relation annotation but some parsers manage to preserve the same level of performance across text genres.
- there is still an important part of work to do for analyzing syntactic phenomena which are rarely or never handled by the actual parsers (apposition or juxtaposition relation, or when coordination are combined together or mixed up with ellipses),
- best performances obtained by different parsers (different performance profiles), so there is *a priori* a relatively important margin for performance increase which could be obtained by combining the annotations of different parsers

First ROVER test

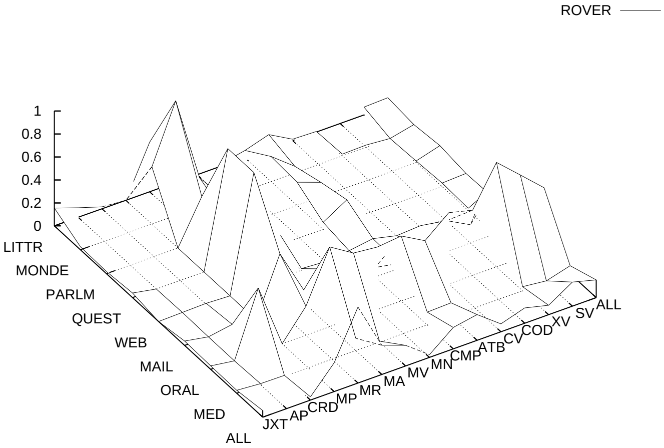

Figure: Relative gain of performance in precision against the best **precision** result

Comparative precision results

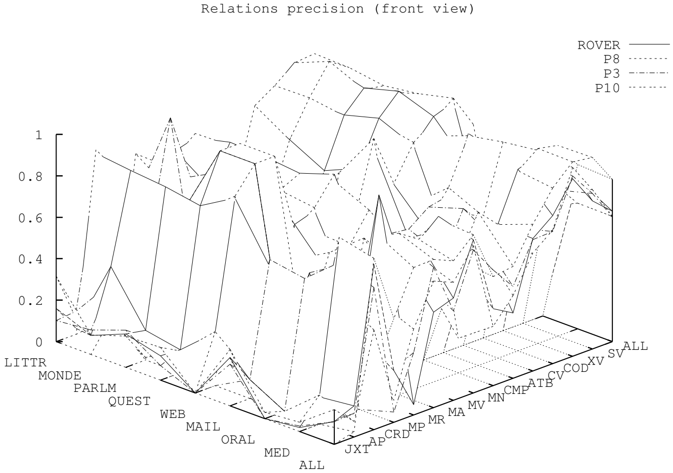

Figure: Compared precisions of the ROVER and the three best systems

Conclusion and perspectives

From EASY to PASSAGE...

- first campaign deploying the evaluation paradigm in real size for syntactic parsers of French with a black-box evaluation scheme using objective quantitative measures.
- create a working group on parsing evaluation
- the beginning of PASSAGE... in a few minutes!