2. Figures

3. Questions

4. Properties of Old French (OF)

5. The SRFMC Corpus

6. The Grammar Model

9.3 Cross Evaluation “Leave One Out” – handout

9.2 Cross Evaluation “Leave One Out”

Philo: I need to parse Old French texts of different types (verse, prose, dialects, etc.) Do I have existing parser models?

Computational Linguist: You won’t lose much if you train the parser on all the data you have. P1: can’t use the same test material. What can I expect from existing parser models?

C: If the training corpus contained 12th century verse texts, you would be prepared for most styles of Old French, including prose — except for the very old-est texts.

P: And what if I want to very old texts?

C: Then the time lapse between your text and the training data should be as small as possible.

P: A golden rule to go home with?

C: Don’t train on prose if you want to parse verse.

P: ADI

Language: Old French

Training Corpus: 200,000 words, 19,800 sentences, many different types (verse, prose, dialects) and different authors.


Parser: mate tools (Bohten, 2010)

http://code.google.com/p/mate-tools/

Scores: 69.6% unlabelled, 82.6% labelled

Why “rich grammar”?

C: 89.68% unlabelled, 82.62% labelled

Parser: A

P: train/eval: all 90/10

Train/eval: 74.93/77.50

Train/eval: 80.36/83.51

Train/eval: 89.68/92.96

Train/eval: 82.68/86.22

Then the time lapse between your text and the parsing purposes. It has a large set of categories and many semantic distinctions.

Table 1: SRFMC texts used here

Table 2: Some SRFMC categories

7. Preparation and Method

• Verified manual dependency annotation and export to CoNLL format using Notabene (Mazzotta, 2010).

• Verified PoS annotation using the Cattaget dataset of the BFM database (http://bfm.ens-lyon.fr)

• Cattaget has part of speech and subcategories, e.g. ADJqua (adjectif qualificatif), but no information on grammatical number.

• Unlabelled Lemmas added by TesTagger (Schmid, 1997) trained on the Nouveau Corpus d’Amsterdam (Kunstmann and Stein, 2007).

The Cattaget dataset did not improve parsing accuracy, but scores for some dependencies of the verb are slightly better.

• Immediate Lemmatization

• Slightly modifications of the grammar model.

• Parsing with mate tools package including Bohten’s graph-based dependency parser (Bohten, 2010). The mates POS/POStagger was trained on the training set using 10-fold jackknifing.

8. Global Results

• 90/10 training evaluation split

• Evaluation on every 10th sentence

(To preserve the heterogeneity of the whole corpus)

jackknifing through ten 90/10 splits, for the complete corpus and a 40,000 word test text

Table 3: 10-fold cross evaluations, mean, standard deviation, CV coefficient of variation

9.3.1 Cross Evaluation “One on One” – handout

The impact of heterogeneity: how does a model trained on one text perform on each of the others?

Roland model: best for the oldest texts. Highest attachment scores and sentence matches in Cattaget (Table/handout: compare horizontally), and better than others for the three other 11th c. texts (Table/handout: compare vertically).

Roland model: best mean values for all scores (Table: column µ)

Lap. model: worst mean values (too small!).

80.36 83.51 70.10 88.53

89.10 92.96 82.32 96.63

84.35 89.68 79.12 90.18

74.98 77.50 62.01 86.75

81.22 82.68 69.95 86.66

82.31 82.78 69.85 85.85

21.36 34.29 17.30 36.21

Table 4: Cross evaluation prose/verse

• Verse training + prose parsing: significantly better results than prose training + verse parsing (differences of 10% for label accuracies and LAS).

• Syntax in verse is more variable.

• A parser trained on verse has seen more variants, and hence performs better.

The relatively good results (compared to the global result in Table 3) obtained on the prose/90 split also indicate that there is less cross syntax.

9.4 General Observations

• High difference between labelled and unlabelled attachment scores (about 10%). The price for our “rich” annotation model? Semantically motivated oppositions between categories are hard to learn for a parser.

• The generally low score of exact matches could indicate that a high number of dependencies are difficult to learn, whereas a small number of dependencies are difficult to learn. The variation of exact matches is about three times higher than the variation of the attachment scores.

10. Conclusions and Future Work

• Cross evaluation results indicate considerable syntactic variation.

• The parser performs surprisingly well even if it is trained globally, i.e. on the complete, heteroge-

neous training corpus.

• The results suggest that the syntactic properties of the text genre should be analysed in greater detail.

References