LIDILEM

UNIVERSITÉ Grenoble Alpes

Annotation of multiword expressions in French

Agnès Tutin, Emmanuelle Esperança-Rodier, Doriane Simonnet, Pauline Soutrenon, Zied Elloumi

LIDILEM/LIG
Université Grenoble Alpes

Aimwest/Parseme-FR Workshop, Grenoble, 3-4 october 2016

Outline

1. **Why shall we annotate MWEs (and why it is not a trivial task)?**
2. **Corpora and annotation scheme**
3. **Semi-automatic annotation of expressions**
4. **Results of annotation**
5. **Some complex cases**
6. **Conclusion and perspectives**

1. Why shall we annotate MWEs in corpora?

- **Theoretical aims :**
  - **To validate a typology of MWEs**
  - **To determine the most frequent MWEs, especially according to specific genres.**
    - **E.g. Are idiomatic metaphoric expressions more frequent in spoken genres?**
    - **E.g. Are collocations more frequent than true idiomatic expressions?**
  - **To observe the syntactic properties of MWEs**
    - **MWEs are highly variable and few of them are « frozen expressions » (Cf Moon 1998)**

Why annotate MWEs in corpora? (2)

- **Practical goals**
  - Few MWE annotated corpora, especially in French
    - **Small corpora with adverbial and nominal MWEs (Laporte *et al.* 2008, Laporte & Voyatzi 2008),**
    - **FrenchTreebank (Abeillé) : 1 million words but few verbs and only contiguous verbs (e.g. *faire part*) and no discontinuous expressions (e.g. *prendre ce problème en compte*).**
    - **Schneider *et al.* 2014's social web corpus with MWE annotations (distinction between strong and weak MWEs)**

    → **No fine-grained typology of MWEs.**

  - Useful for MT applications to evaluate which MWEs are more difficult to translate
    - **Hypothesis (partially) confirmed by a first LIG-LIDILEM study: contiguous MWEs are easier to translate**

Why annotate MWEs in corpora? (3)

- **But this is not a trivial task:**

1. Is an expression a MWE?
   - **Easy for compounds (*as long as*) and full phrasemes (*to spill the beans*), complex for collocations or routines**
2. Delimiting the boundaries of the expression
   - **Include or not determiners in verbal MWEs?**
     - In our annotation scheme, inclusion of fixed determiners, omission of variable determiners

       il fait la fête

       But *elle donne un cours*
   - **Include or not the auxilary for verbs?**
3. Which kind of MWEs?
   - **Collocation? Full Phraseme? Term? Pragmateme?**

2. Corpora and annotation scheme

- **Aim :**
  - **Creation of an MWE annotated corpus of 69,000 tokens**
  - **A varied bilingual corpus, freely available, on different genres (French texts annotated only):**
    - Scientific writing : BAF Citi 1 (Baf corpus) : ~ 14,500 tokens
    - News (journalese) : ~ 17 ,400 tokens
    - Subtitles of *Amélie Poulain* : 9,900 tokens
    - Excerpt of *Thérèse Raquin* (Zola) : 7,260 tokens
    - TED talks : 8,160 tokens (but poor quality of the translation)
    - Oral transcriptions EIIDA of scientific talks (monolingual) : 12,630 tokens
  - **Several types of MWEs**
  - **Semi-automatic annotation of the corpus with finite-state tool (NooJ system, Silberztein *et al.* 2013)**

Annotation scheme
Typology of MWEs

- **Inspired by Granger & Paquot (2008), Heid (2008), Tutin (2010), Mel'čuk (2011)**
  - **« Full phrasemes » (non compositional)**
    - **Nominal, adjectival, adverbial compounds and verbal phrasemes :** *pomme de terre* ('potatoe '), *dead end, bon marché* ('cheap'), *to take into account*
  - **Collocations or semi-phrasemes (including light verb constructions)**
    - *To have a shower, heavy smoker* vs. *gros fumeur, freshly baked*
  - **Functional MWEs**
    - **Functional adverbs, prepositions, conjunctions , determiners, pronouns:**
      on the one hand, in front of, insofar as, a large number of
  - **Pragmatemes (spoken)**
    - *You're welcome, see you later*

Annotation scheme
Typology of MWEs (2)

- **Proverbs**
  - *Jack of all trade, master of none. First come, first served*
- **Complex terms**
  - *Natural language processing, syntactic parser*
- **Named entities**
  - *Université Stendhal, Laboratoire d'Informatique de Grenoble*
- **Routine formulae**
  - *As previously said, force est de constater …*
- **Phrasal verbs (for Germanic langages)**
  - *Give up*

Annotation scheme

- **Principles : simple surface annotation**
  - **Stand-off annotation: better suited but too complex for linguists**
- **Features**
  - **Identifier**
  - **Type of MWE : full phraseme, collocation, complex term …**
  - **Syntactic category of full expression : verb, adverb, noun, …**
  - **Syntactic category of each part of the MWE**

Example

Nous avons pris ce problème en compte (‘we hake taken into account this problem’).

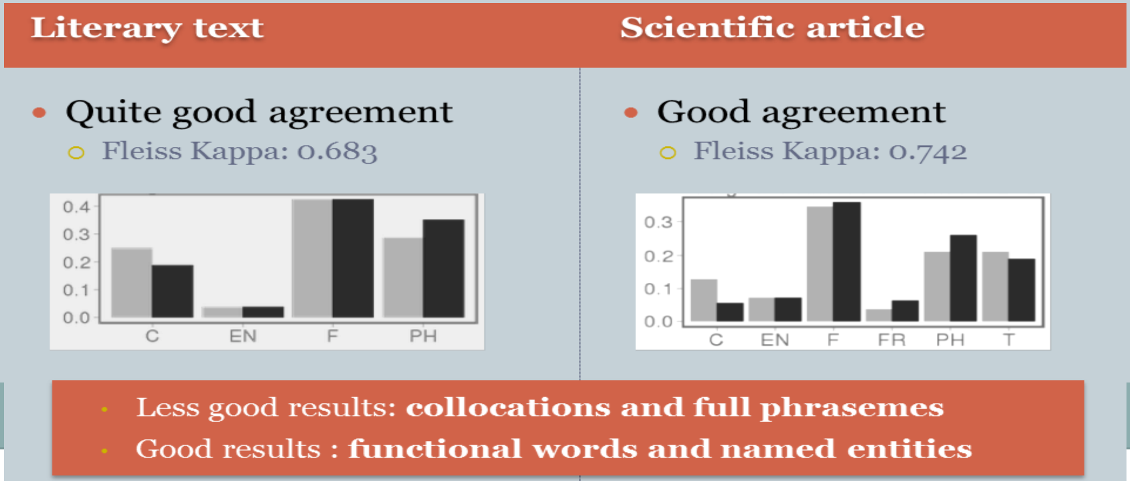
```
Nous avons <epl num="23" type="PH" mcat="V"
pos="V">pris</epl> ce problème <epl num="23" type="PH"
mcat="V" pos="P">en</epl> <epl num="23" type="PH"
mcat="V" pos="P" pos="N">compte</epl>
```

Annotation scheme

- **Overlapping expressions are possible**
  - **Partial overlapping : *pay attention* + *close attention***
    - Unlike many theorists, he **[paid$_1$] [close$_2$] [attention$_{1+2}$]** to a broad range of experimental evidence…
  - **Inclusion : *au minimum* included in the collocation *réduire au minimum* (‘reduce to a minimum’)**
    - Afin de **[réduire$_1$] [au$_{1+2}$] [minimum$_{1+2}$]** cet effort …

3. Annotation process

- **Semi-automatic annotation process:**
  - Using a lexicon of MWEs
  - Surface annotation with finite state techniques (Nooj, Silbertzein *et al.* 2013)
  - Double annotation (at least, two annotators) and automatic checking of consistency (detection of identical MWEs, etc.)
  - Reasonable interannotator agreement (Tutin *et al.* 2016)
    - Calculated on the type of MWE, on MWEs annotated by both annotators, on an extract of 4000 tokens in scientific report and novel



Literary text

- Quite good agreement
  - Fleiss Kappa: 0.683

Scientific article

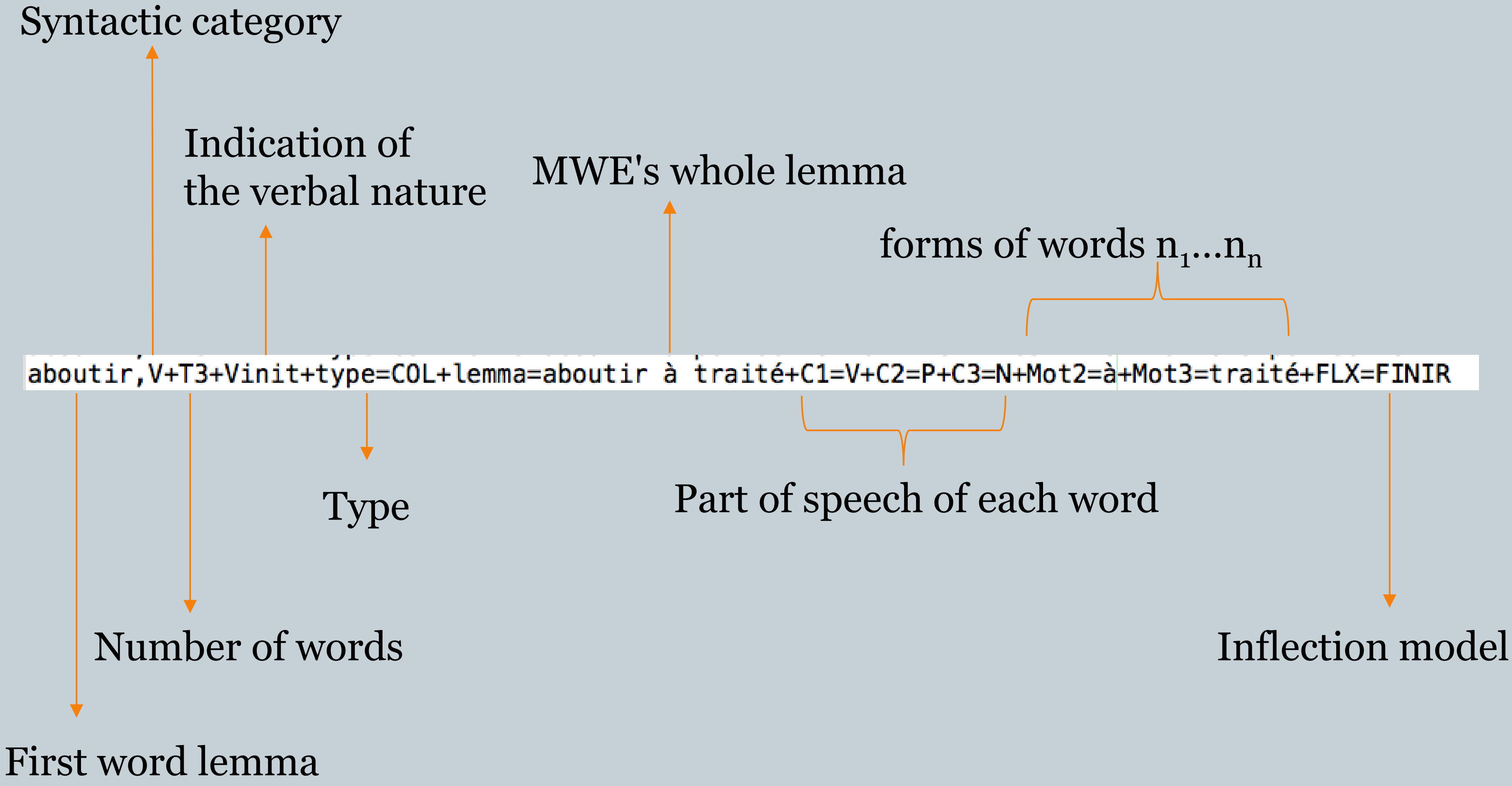
- Good agreement
  - Fleiss Kappa: 0.742

- Less good results: **collocations and full phrasemes**
- Good results : **functional words and named entities**

3. Annotation process

- **Dictionary used: a core lexicon collected from several sources**
  - Extracted MWEs from FrenchTreeBank (Abeillé *et al.* 2003)
    - **Interest : most frequent MWEs, decomposition of MWEs**
  - A subset of several dictionaries:
    - **Dictionnaire Electronique des Mots (Dubois et Dubois Charlier)**
    - **Wiktionary for French**
    - **DELAC (Courtois *et al.* 1997)**
    - **Fontenelle's (1992) collocations database (extracted from Robert and Collins Dictionary)**

Example of semi-automatic annotation with NooJ (FST)



Syntactic category

Indication of
the verbal nature

MWE's whole lemma

forms of words $n_1 \ldots n_n$

```
aboutir,V+T3+Vinit+type=COL+lemma=aboutir à traité+C1=V+C2=P+C3=N+Mot2=à+Mot3=traité+FLX=FINIR
```

Type

Part of speech of each word

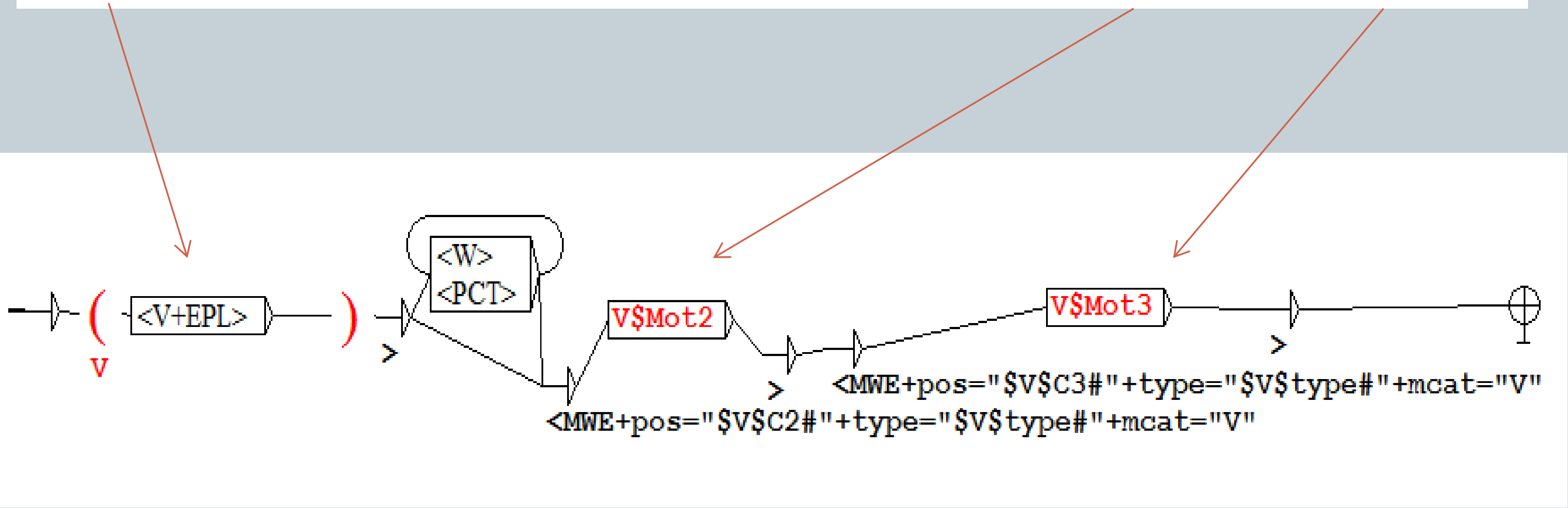
Number of words

Inflection model

First word lemma

Example of semi-automatic annotation with NooJ (FST)

faire,V+EPL+type=PH+lemma=faire les frais+C1=V+FLX=FAIRE+C2=D+Mot2=les+C3=N+Mot3=frais

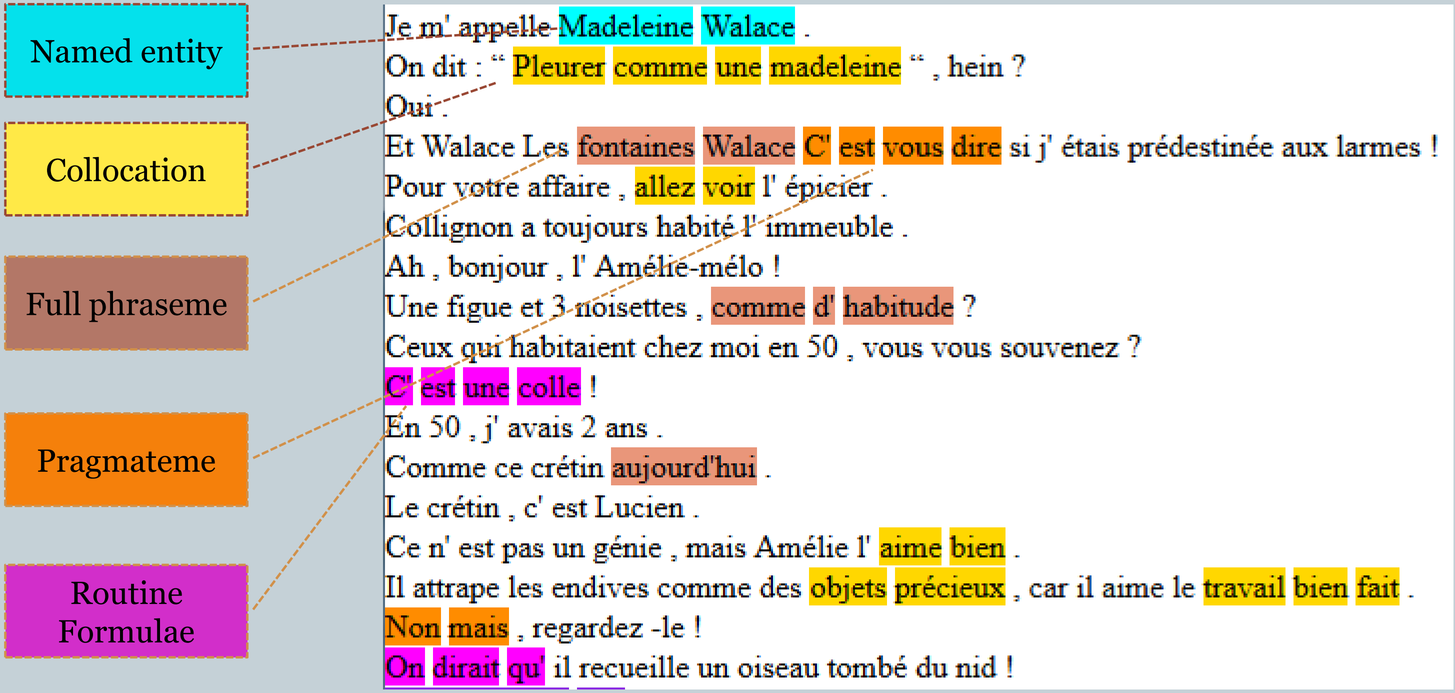

(
V
<V+EPL>
)
<W>
<PCT>
V$Mot2
V$Mot3
<MWE+pos="VC2#"+type="Vtype#"+mcat="V"
<MWE+pos="VC3#"+type="Vtype#"+mcat="V"
>

Annotation process

- **MWE lexicon of about 5000 MWEs with syntactic decomposition and a MWE type**
- **Semi-automatic annotation performed and checked on concordances, and completed with manual annotation with the help of an XML editor (Oxygen, with annotation guidelines)**
- **About 35% to 50% of MWEs are semi-automatically annotated**
  - **Good coverage of functional words and frequent phrasemes**
  - **Weak coverage of collocations, pragmatemes , terms and routines**
    - **Could be easily improved with a better named entity recognition system**
    - **Complex terms are difficult to detect : => use of a term detector?**

Example of annotated text

- ## Subtitles of the film *Amélie Poulain*



Named entity

Collocation

Full phraseme

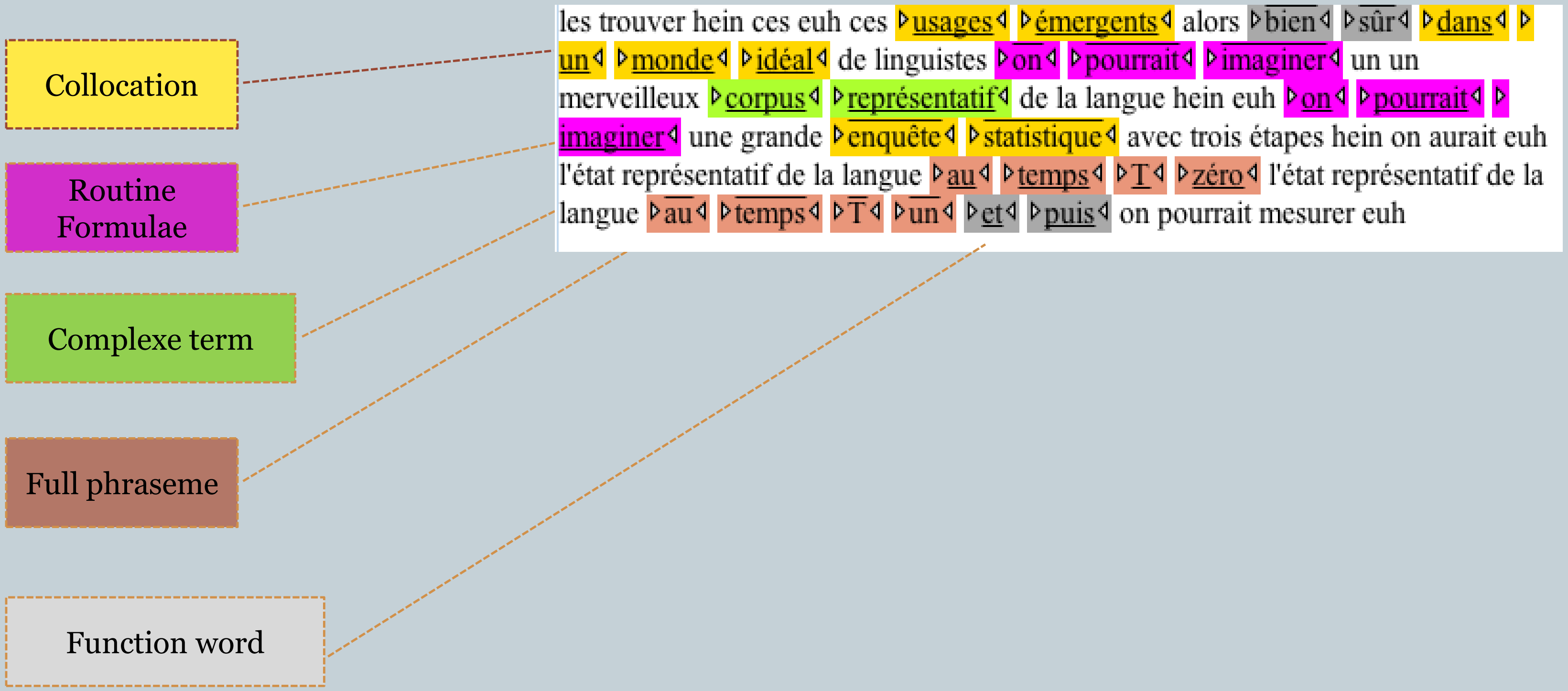
Pragmateme

Routine Formulae

Je m' appelle Madeleine Walace .
On dit : " Pleurer comme une madeleine " , hein ?
Oui .
Et Walace Les fontaines Walace C' est vous dire si j' étais prédestinée aux larmes !
Pour votre affaire , allez voir l' épicier .
Collignon a toujours habité l' immeuble .
Ah , bonjour , l' Amélie-mélo !
Une figue et 3 noisettes , comme d' habitude ?
Ceux qui habitaient chez moi en 50 , vous vous souvenez ?
C' est une colle !
En 50 , j' avais 2 ans .
Comme ce crétin aujourd'hui .
Le crétin , c' est Lucien .
Ce n' est pas un génie , mais Amélie l' aime bien .
Il attrape les endives comme des objets précieux , car il aime le travail bien fait .
Non mais , regardez -le !
On dirait qu' il recueille un oiseau tombé du nid !

Example of annotated text

- **Oral transcription EIIDA (scientific talk)**



les trouver hein ces euh ces ▷usages◁ ▷émergents◁ alors ▷bien◁ ▷sûr◁ ▷dans◁ ▷un◁ ▷monde◁ ▷idéal◁ de linguistes ▷on◁ ▷pourrait◁ ▷imaginer◁ un un merveilleux ▷corpus◁ ▷représentatif◁ de la langue hein euh ▷on◁ ▷pourrait◁ ▷imaginer◁ une grande ▷enquête◁ ▷statistique◁ avec trois étapes hein on aurait euh l'état représentatif de la langue ▷au◁ ▷temps◁ ▷T◁ ▷zéro◁ l'état représentatif de la langue ▷au◁ ▷temps◁ ▷T◁ ▷un◁ ▷et◁ ▷puis◁ on pourrait mesurer euh

4. Results of annotation

- Phraseological behaviour according to the annotated texts' genres

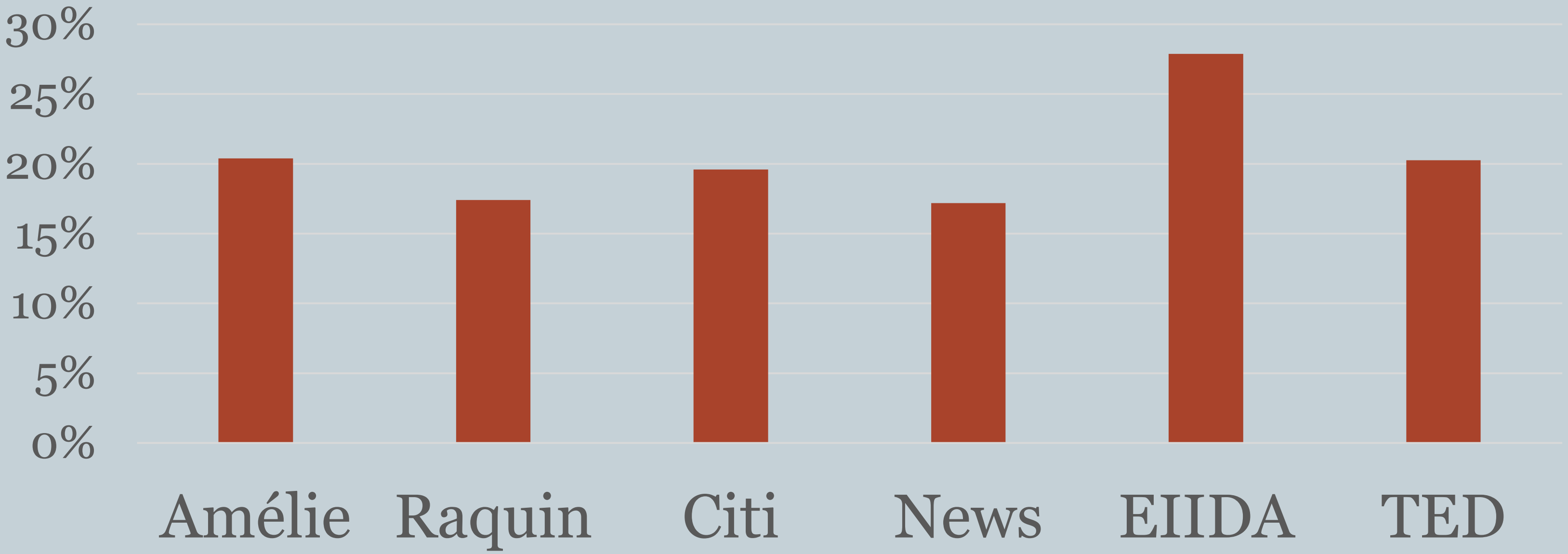
Spotting MWEs in the annotated texts

- ○ Observation of the phraseological density
- ○ Observation of the phraseological types

💣 Small corpus

Phraseological denstity

- % of the « tokens » included in the MWEs.
- Important proportion of the phraseology (1/6 to 1/5 of the tokens), quite stable from one genre to another
- Similar to *Schneider et al.* (2014) on English



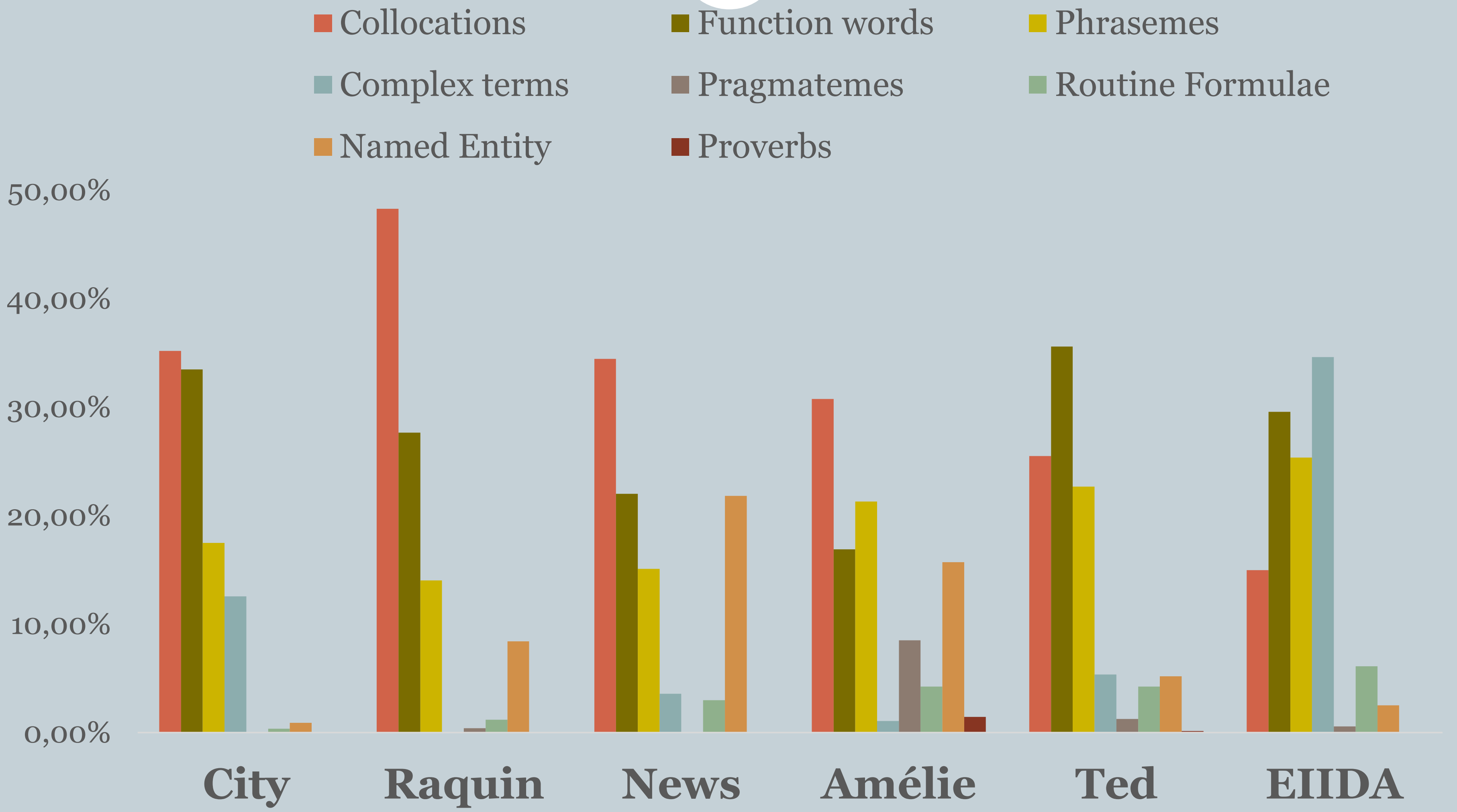
Phraseological density

- Few differences between the genres
- Oral transcription contains slightly more MWEs than the other genres
    - To be confirmed on a larger corpus?

Type distribution

- MWEs distribution
  - Hypothesis : important variation of some MWEs' types among the genres
  - Stable types?

Types distribution



Types distribution

- Some types are specific to one genre
  - Complex terms → Scientific writings
  - Pragmatemes et proverbs → Dialogues
- Three frequent types constantly appearing in all genres
  - Collocations
  - Function words
  - Full phrasemes
- Some types are less frequent
  - Routine formulae

5. Some complex cases

- MWEs are often ambiguous
  - *"il y a" : there is = Function words / [time] ago = Phraseme*
- Distinction between types is not trivial
  - *Crise cardiaque* : full phraseme or complex term ?
- Routine formulae
  - *Pour résumer…, Le problème est le suivant…*
  - *Je pense…, Croyons-nous…* (close to pragmatemes)
  - Non par […] mais par…, Pas tant […] que… (syntactic schemes)
  - Dans les années…[NUM], En plein [N] (lexical patterns)

Some complex cases

- Annotation of determiners in verbal collocations
  - Fixed determiners are annotated
  - Annotate the variability?
  - *Donner un cours*
  - *Se hisser au pouvoir*

Conclusion

- **Annotation of MWEs: a stimulating and feasible task but a complex task for some categories of MWEs, especially in the literary corpus**
  - Need to provide better definitions with formal criteria especially for collocations and phrasemes
  - More detailed examples of compositionality

- **Need for a double annotation (and more in case of disagreement)**
  - To confront interpretations and refine the criteria
- **Automatic annotation needs to be developed**
  - Can be developed incrementally with annotated corpora

Conclusion

- **Annotation of MWEs: a stimulating and feasible task but a complex task for some categories of MWEs, especially in the literary corpus**
  - Need to provide better definitions with formal criteria especially for collocations and phrasemes
  - More detailed examples of compositionality
- **Need for a double annotation (and more in case of disagreement)**
  - To confront interpretations and refine the criteria
- **Automatic annotation needs to be developed**
  - Can be developed incrementally with annotated corpora

Perspectives

- An experiment of SMT evaluation (translation of MWEs) on a subset of the corpus
- Lexical alignment of the MWEs for MT applications and pedagogical applications
- Participation to the Parseme –FR annotation shared task (with the news corpus)?

Thank your for your attention (pragmateme?)

References

- ABEILLÉ, A., CLÉMENT, L. AND L. TOUSSENEL, 2003. Building a treebank for French. In: Treebanks. Springer Netherlands. pp. 165-187.
- COURTOIS, B., GARRIGUES, M., GROSS, G., GROSS, M., JUNG, M., MATHIEU-COLAS, M., MONCEAUX, A., PONCET-MONTANGE, SILBERZTEIN, M. AND VIVÈS, R., 1997. Dictionnaire électronique DELAC : les mots composés binaires. Technical Report 56, University Paris 7, LADL.
- DUBOIS, J. AND DUBOIS-CHARLIER, F. 2010. La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages*, 179(3, pp 31-56.
- FONTENELLE, TH. 1997. Turning a Bilingual Dictionary into a Lexical-Semantic Database Tübingen: Max Niemeyer Verlag.
- GRANGER, S. AND PAQUOT, M. 2008. Disentangling the phraseological web. In Granger, S. & Meunier, F. *Phraseology: An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins.
- HEID, U. 2008. Computational phraseology. An overview. In: S. Granger and F. Meunier, *Phraseology*. *An interdisciplinary perspective*. Amsterdam: Benjamins. pp 337-360.
- LAPORTE, E., NAKAMURA, T., AND VOYATZI, S. 2008a. A French corpus annotated for multiword nouns. *In Language Resources and Evaluation Conference*. Workshop Towards a Shared Task on Multiword Expressions. pp. 27-30.
- LAPORTE, E., NAKAMURA, T., AND VOYATZI,. 2008b. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC)*. Linguistic Annotation Workshop. pp. 48-51.

References

- MEL'ČUK, I. 2013. Tout ce que nous voulions savoir sur les phrasèmes, mais… *Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie*, 102, pp. 129-149.
- MOON, R. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford:Oxford University Press.
- POTET, M., ESPERANÇA-RODIER, E., BESACIER, L., BLANCHON, H. 2012. Collection of a Large Database of French-English SMT Output Corrections, *(LREC 2012)*, Istanbul, 2012, 21-27 mai.
- SCHNEIDER, N., ONUFFER, S., KAZOUR, N., DANCHIK, E., MORDOWANEC, M. T., CONRAD, H., AND SMITH, N. A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. *Proc. of LREC*. Reykjavík, Iceland.
- SILBERZTEIN, M., KHURSHUDIAN, V., AND DONABÉDIAN, A., 2013. *Formalizing Natural Languages with Noo*. Cambridge: Cambridge Scholar Press.
- TUTIN, A. 2010. *Sens et combinatoire lexicale: de la langue au discours*. Unpublished Dossier en vue de l'habilitation à diriger des recherches). Grenoble: Université Stendhal.
- TUTIN, A., ESPERANÇA-RODIER, E., IBORRA, M., REVERDY, J. 2016. Annotation of multiword expressions in French. Corpas-Pastor Gloria. *European Society of Phraseology Conference (EUROPHRAS 2015)*, Jun 2015, Malaga, Spain. pp.60-67,