Recent Work at LIG on the AIM-WEST Project

Laurent Besacier, Zied Elloumi, Agnès Tutin, Emmanuelle Esperança-Rodier

LIG
Université Grenoble Alpes

AIM-WEST Meeting, Porto-Alegre,
20-21 oct 2015

Outline

1. **Recall on 2014 work**
2. **MWE Translation shared task for French-English language pair**
3. **Annotation of multiword expressions in French**
4. **Proposition of a new MT evaluation metric based on LIG semantic resource (DBnary)**

Associated Papers

1. Zied Elloumi, Olivier Kraif, Laurent Besacier. "**Integrating Multi Word Expressions in Statistical Machine Translaton**", Proceedings of Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT2015) – EUROPHRAS2015 – Malaga, Spain, 1-2 July **2015**.
2. Agnès Tutin, Emmanuelle Esperança-Rodier, Manolo Iborra, Justine Reverdy "**Annotation of multiword expressions in French** " Proceedings of Conference EUROPHRAS2015 – Malaga, Spain, 1-2 July **2015**.
3. Zied Elloumi, Hervé Blanchon, Gilles Sérasset and Laurent Besacier, "**METEOR for Multiple Target Languages using DBnary**", Proc. 15th MT Summit 2015, Oral, Miami, Florida, USA, 30Oct-3Nov **2015**

Outline

1. **Recall on 2014 work**
2. **MWE Translation shared task for French-English language pair**
3. **Annotation of multiword expressions in French**
4. **Proposition of a new MT evaluation metric based on LIG semantic resource (DBnary)**

Recall of 2014

- **English-French MWE Translation**
  - Focus on idioms and phrasal verbs
  - Pre-processing of the training / development data
    - **He picked up a remote and turned the stereo on**

      => He picked up a remote and turned_on the stereo
    - **And you thought my job was <idiom translation="facile">a piece of cake</idiom>**
  - SMT Performance improvements on our « challenging » test corpus for MWE translation
    - **Z. Elloumi, O. Kraif, L. Besacier. *"Integrating Multi Word Expressions in Statistical Machine Translaton"*, Proceedings of MUMTTT2015, Malaga, Spain, 1-2 July 2015.**

In 2015…

- **French-English MWE Translation**
  - Proposed as a « shared-task » to AIM-WEST partners
  - Organized by LIG
    - Training/Test material made available by Zied
    - Systems Error analysis was scheduled

    http://aim-west.imag.fr/mwe-translation-shared-task-is-now-launched/
  - But … only LIG submitted runs …
    - Similar approach compared to 2014…
    - …but for French-English SMT
    - Focus on collocations and idioms

Fr-En MWE Translation

Idiom	
Source	Je ne veux jeter ici la pierre à personne
Reference	I do not want to start apportioning blame now
Google Translate	I do not want to throw stones at anyone here

Fr-En MWE Translation

Collocation	
Source	J' espère que ce débat sur la communication de la commission **donnera lieu** enfin à une action européenne efficace .
Reference	I hope that this debate on the commission communication will finally **lead** to effective european action in this field .
Hyp Moses-LIG	I hope that this debate on the communication from the commission to finally **give place** for effective european action .

« Challenging » Test Corpus

- **Using LIDILEM's tool - Lexico-gramme [O. Kraif et S. Diwersy, 2012]**
- **3 types of collocations**
  - **Support verb constructions**
  - **VERB + NOUN [object] : forcer admiration, annoncer nouvelle**
  - **NOUN [subject] + VERB : le ton monte, le temps d'écoule**
- **Idioms**

	Corpus FR-EN	Collocations	Idioms
Literary Texts	206	160	46
Europarl	290	261	29
News	4	4	0
TOTAL	**500**	**425**	**75**

Google *vs* Moses

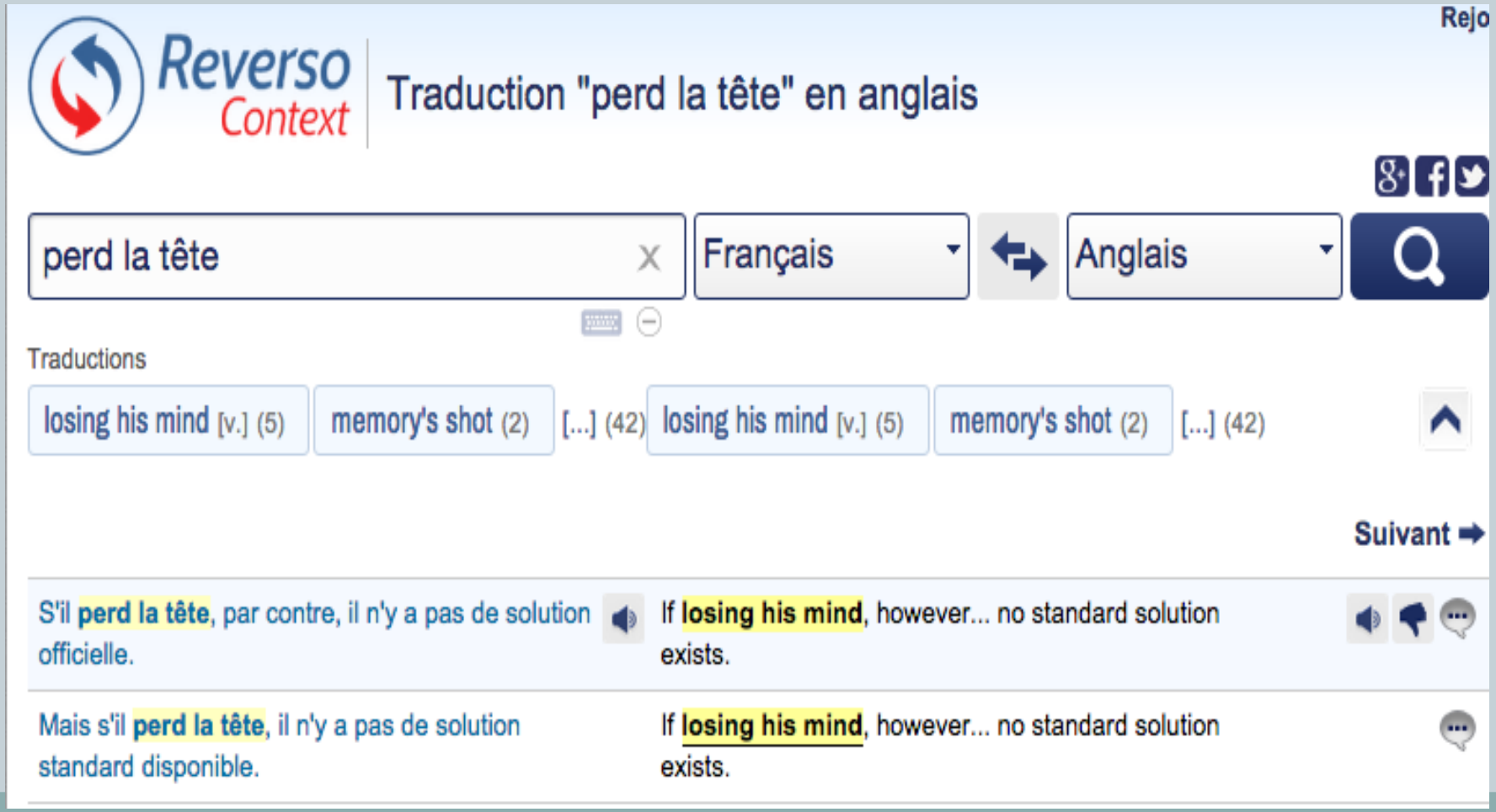
- **Phrase-Based Fr-En SMT using Moses (Koehn, 2007)**
- **Training Data : Europarl, TED, News**
- **Extracting a *Dev* corpus with the same distribution (Litt., Europarl, News) as our *Test* Corpus**
- **BLEU metric (Papineni and al. 2002) for evaluation**
- **Using MERT (Och and Ney, 2003) for SMT tuning**

	Dev	Test (MWE)	Collocations	Idioms
Moses-LIG	**29.62**	26.05	26.17	25.15
Google-Tr	26.21	**28.90**	**28.25**	**30.89**

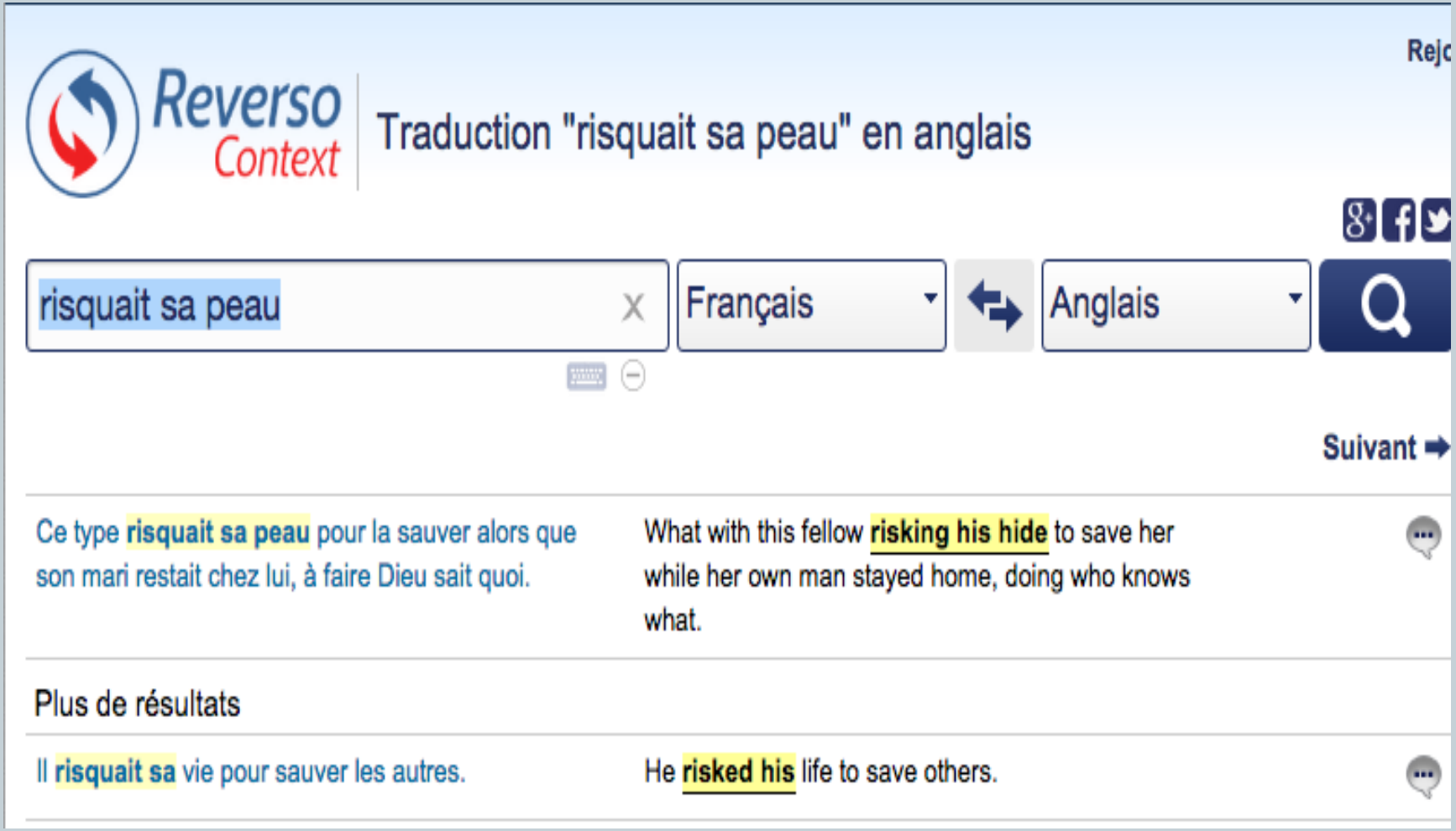
Handling MWE Translation

- **Pre-processing of the training / development /test data**
  - **Le garçon accusa le coup**
    => Le garçon accusa-le-coup
- **Idioms**
  - **Pre-defined list (from *Test*) using Reverso Context**
  - **http://context.reverso.net/traduction/**
  - **<EPL translation="giving voice" > donner de la voix </EPL>**
  - **<EPL translation="losing his mind"> perd la tête </EPL>**

Reverso Context



Reverso Context



SMT Results

	Idioms 75 sent	TEST-collocations (207 sent)	TEST-collocations +Idioms
Baseline	24.17	29.39	28.12
Our approach	**26.13**	**30.13**	**29.16**

Baseline I hope that this debate on the communication from the commission to finally ***give place*** for effective european action .

MWE Specific I hope that this debate on the commission 's communication will **lead** finally to effective european action .

SMT Results

	Idioms 75 sent	TEST-collocations (207 sent)	TEST-collocations +Idioms	TEST-MWE (425 sent)	TEST-MWE +Idioms (500 sent)
Baseline	24.17	29.39	28.12	**26.17**	**26.05**
Our approach	**26.13**	**30.13**	**29.16**	25.04	25.20

Baseline I hope that this debate on the communication from the commission to finally ***give place*** for effective european action .

MWE Specific I hope that this debate on the commission 's communication will **lead** finally to effective european action .

Outline

1. **Recall on 2014 work**
2. **MWE Translation shared task for French-English language pair**
3. **Annotation of multiword expressions in French**
4. **Proposition of a new MT evaluation metric based on LIG semantic resource (DBnary)**

Why 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?

- **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?

- **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**
1. **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 <u>fait</u> <u>la fête</u>

       But *elle <u>donne</u> un <u>cours</u>*

   - **Include or not the auxilary for verbs?**
1. **Which kind of MWEs?**
   - **Collocation? Phraseme? Term? Pragmateme?**

Corpora and annotation scheme

- **Aim**
  - **Creation of an MWE annotated corpus of 45,000 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 : ~ 12 ,300 tokens
    - Subtitles of *Amélie Poulain* : 9,900 tokens
    - Excerpt of *Thérèse Raquin* (Zola) : 7,260 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

- **Proverbs**
  - *Jack of all trade, master of none. First come, first served*
- **Multiword 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**

- **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’).

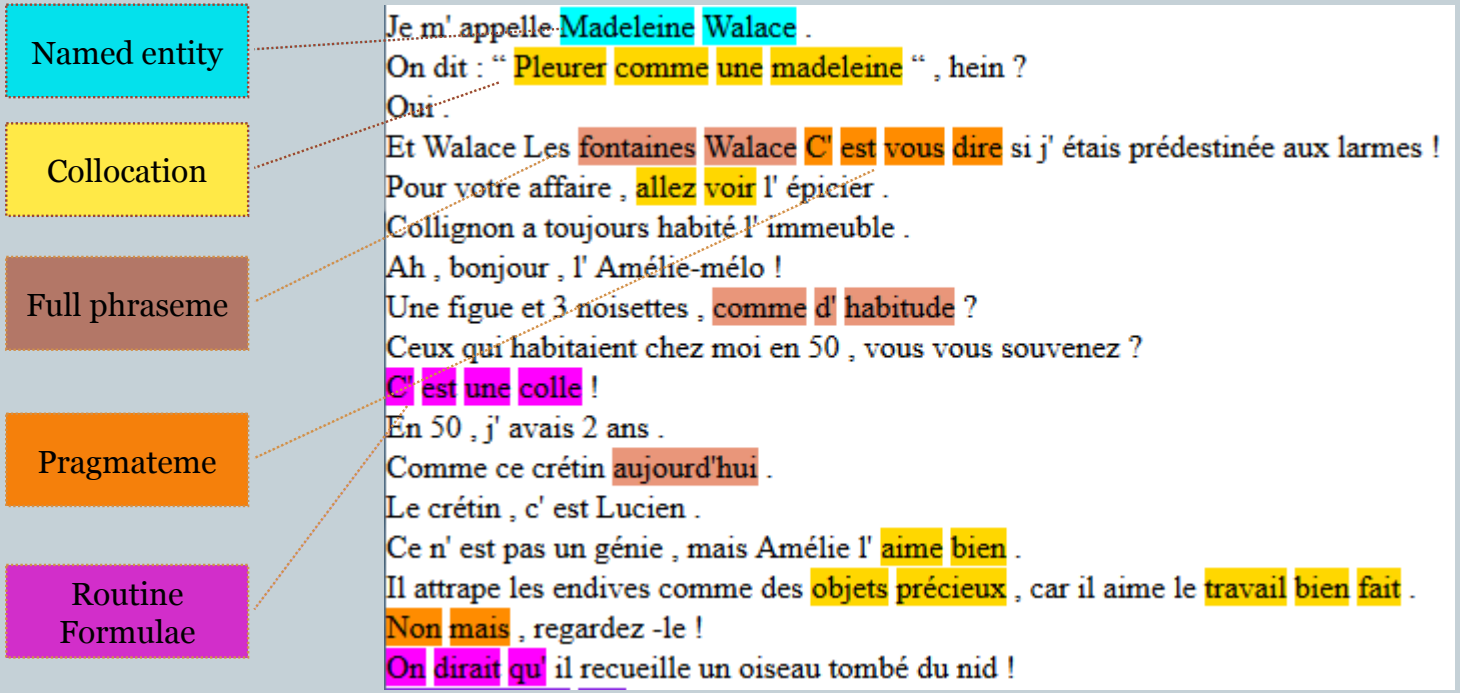
```
Nous avons <epl id="23" type="fphraseme" catepl="verb"
catw="verb" lemma="prendre_en_compte">pris</epl> ce
problème <epl id="23" catw="prep">en</epl> <epl id="23"
catw="noun">compte</epl>
```

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 …

Example of annotated text

- **Subtitles of the film *Amélie Poulain***

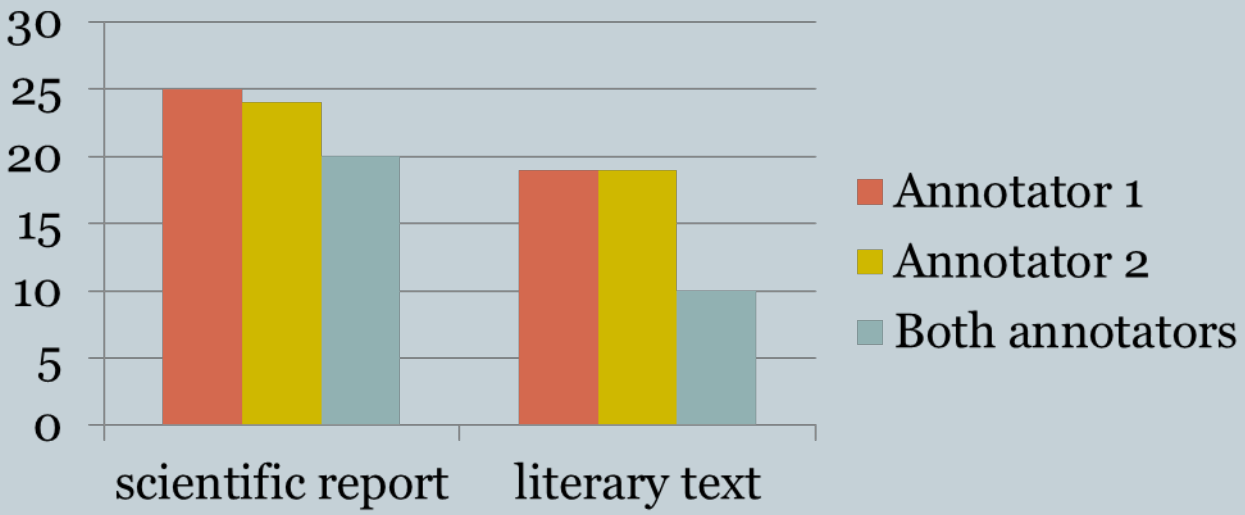


Results

- **Evaluation on two extracts of our corpus**
  - Corpus :
    - **Literary text (*Thérèse Raquin*) : ~ 2000 words**
    - **Scientific report (*CITI 1*): ~ 2000 words**
  - Two annotators :
    - **Linguists familiar with the issue of MWEs**
  - Grammatical issues are not taken into account

Quantitative results: proportion of MWEs

- **% of words belonging to a MWE**

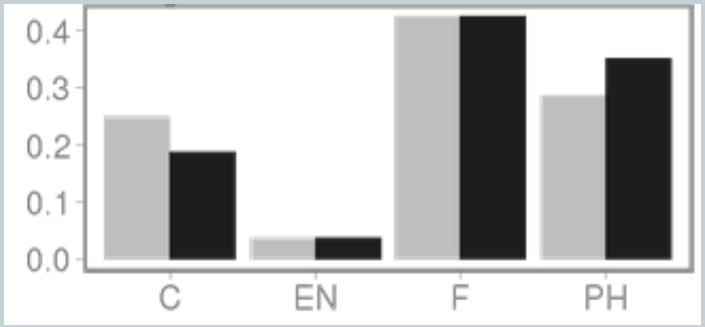


  - More MWEs in the scientific text (due to formulaic language?)
  - Better agreement on what is a MWE in the scientific text (see next slide)

Quantitative results: agreement on the type of MWE

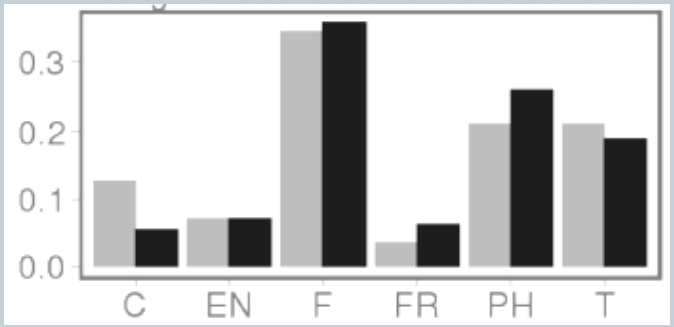
Literary text

- 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**

Conclusion

- **Annotation of MWEs: a stimulating and feasible task but a complex task for some categories of MWEs, especially in the literary corpus**
- **Need for a double annotation (and more in case of disagreement)**
  - **To confront interpretations and refine the criteria**
- **(Semi-) automatic annotation needs to be developed**
  - **Can be developed incrementally with annotated corpora**
- **Annotation guidelines available online**
  http://aim-west.imag.fr/wp-content/uploads/2014/04/Annotation-guidelines-for-MWE_20_april_2015.pdf

Outline

1. **Recall on 2014 work**
2. **MWE Translation shared task for French-English language pair**
3. **Annotation of multiword expressions in French**
4. **Proposition of a new MT evaluation metric based on LIG semantic resource (Dbnary)**

   (separate presentation)

References

- Kraif, O. and Diwersy, S.. 2012. *Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'ex- traction de constructions lexico-syntaxique*s, Actes de la conférence TALN 2012, Grenoble France: 399–406.

- Och, F. 2003. *Minimum Error Rate Training in Statis tical Machine Translatio*n, ACL '03 Proceedings of the 41st Annual Meeting on Association for Com- putational Linguistics, Vol 1, USA:160–167.

- Papineni, K., Roukos, S., Ward, T., and Zhu W. 2002. *BLEU: a Method for Automatic Evaluation of Ma- chine Translation*, Proceedings of the 40th Annual Meeting of the Association for Computational Lin- guistics (ACL), p., July 2002, Philadelphia:311– 318.

- Elloumi, Z., Besacier, L., Kraif, O., 2015. *Integrating Multi Word Expressions in Statistical Machine Translation*, Proceedings of MUMTTT Workshop at EUROPHRAS-15, Malaga-Spain.

- Koehn, P., 2014. *Statistical Machine Translation, System User Manual and Code Guide,* Université of Edinburgh Royaume-Uni.

Thank your for your attention (pragmateme?)