Project in the framework of the AIM-WEST project
Annotation of MWEs for translation

Agnès Tutin
LIDILEM/LIG
Université Grenoble Alpes

30 october 2014

Outline

- **Why annotate MWEs in corpora?**
- **A first experiment**
- **Typology of MWEs**
- **Annotation scheme**
- **Corpora**
- **Annotation process**

Why annotate MWEs in corpora?

- **Theoretical aims :**
  - To validate MWE typology
  - To explore the most productive MWEs according to genres.
    - **E.g. Are really idiomatic metaphoric expressions more frequent in spoken genres?**
    - **E.g. Are really collocations more frequent than true idiomatic expressions?**
  - To observe the syntactic properties of MWEs
    - **Hypothesis : MWEs are highly variables and few of them are « frozen expressions » (Cf Moon 1998)**

Why annotate MWEs in corpora (2)?

- **Practical aims**
  - 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*).**

    ➔ No typology of MWEs.
  - Useful for MT applications to evaluate which MWEs are more difficult to translate
    - **Hypothesis (partially) confirmed by the first LIG-LIDILEM study (internship) : contiguous MWEs are easier to translate**

Why annotate MWEs in corpora?

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

A first experiment in Grenoble

- **Internship (LIF & LIDILEM) (Master students Justine Rverdy and Manolo Iborra, may-july 2014, supervisors : L. Besacier, A. Tutin)**
  - Creation of an MWE annotated corpus of 12500 words (French version of the BAF corpus, scientific corpus)
  - Manual annotation of the corpus with several types of MWEs : collocations, coupounds, functional words, multiword terms, named entities
  - Example of annotation :

‹epl type="Locution verbale" id="14577575" d="0">**être**</epl> <epl type="Locution verbale" id="14577575" d="0">**en**</epl> <epl type="Locution verbale" id="14577575" d="0">**mesure**</epl>

A first experiment (2)

- **Evaluation of the translation of MWEs (Moses)**
  - 80% of translated MWEs are MWEs
  - 70% MWEs are correctly translated (manual evaluation with 3 values : OK, to revise, incorrect)

Type of MWE	Number of occurrences	MWE in target language	Correct Translation	Translation to revise	Incorrect translation
Functional words	390	71%	77%	6%	17%
Nominal and adjectival compounds	72	72%	46%	4%	50%
Complex terms	122	100%	73%	9%	18%
Verbal phrasemes	66	73%	61%	12%	27%
Collocations	124	93%	65%	6%	29%
Total	785	80%	70%	7%	23%

- As expected, most MWEs are translated by MWEs
  - Less MWEs : functional words
  - More MWEs : complex terms and collocations
- As expected, best translations with functional words, and complex terms.
- Unexpectedly, better translations with verbal phrasemes than with adjectival and nominal compounds;

An experiment to be extended?

- **An interesting experiment, worth to be extended**
- **With a larger and more diverse corpus**
  - **Including spoken and written language**
  - **With several genres**
    - **TED corpus**
    - **Literary texts**
    - **Europarl**
    - **Scientific writings**
- **With a semi-automatic annotation process (and a more detailed annotation scheme)**
- **With a more precise evaluation protocol (in collaboration with Emmanuelle Esperança-Rodier)**

Typology of MWEs

- **Inspired by Heid (2008), Mel'čuk (2011),**
  - **« full phrasemes » (non compositional)**
    - **Nominal compounds : *pomme de terre* ('potatoe '), *dead end***
    - **Adjectival compounds : *bon marché* ('cheap')**
    - **Verbal phrasemes : *to take into account***
  - **Functional MWEs**
    - **Adverbs : *on the one hand***
    - **Prepositions : *in front of***
    - **Conjunctions : *insofar as***
    - **Determiners: *a large number of***
  - **Collocations or semi-phrasemes (including lihgt verb constructions)**
    - ***To have a shower, heavy smoker, freshly baked***

Typology of MWEs (2)

- **Pragmatemes (spoken)**
  - ***You're welcome, see you later***
- **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***

Annotation scheme

- **To develop**
- **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
  - Lemma of the expression

Example
Nous avons pris ce problème en compte.

```
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>
```

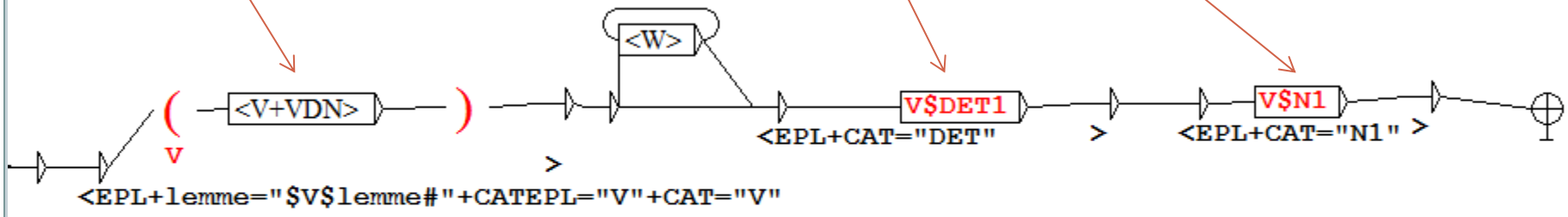
Annotation process

- **Semi-automatic annotation process:**
  - Using MWE lexicons
  - Surface annotation with finite state techniques in a first step, then syntactic parsing (XIP?)

- **Dictionaries used**
  - Extracted MWEs from FrenchTreeBank
    - **Interest : most frequent MWEs, decomposition of MWEs**
  - Dictionnaire Electronique des Mots (Dubois et Dubois Charlier)
    - **Wide coverage, semantic features**
  - Wiktionary
    - **Wide coverage**
  - DELAC
    - **Wide coverage**

Example of semi-automatic annotation with NooJ (FST)

abandonner,V+VDN+lemma=abandonner_la_partie+NoHum+DET1=la+N1=partie+Passif+FLX=AIMER



Grammar to analyse expressions such as "il a abandonné la partie" ou "il a abandonné les lieux"

Manual check of the annotation on concordances and generation of the annotation

Concordance for Text epl.not [Modified]

Reset Display: 5 characters word forms before, and 5 after. Display: Matches Outputs

ext	Before	Seq.
	Max a	abandonné la partie/<EPL+lemme="abandonner_la_partie"+CATEPL="V"+CAT="V""><EPL+CAT="DET"><EPL+CAT="N1">
abandonné la partie. il a		abandonné les lieux/<EPL+lemme="abandonner_les_lieux"+CATEPL="V"+CAT="V""><EPL+CAT="DET"><EPL+CAT="N1">

epl.not.xml.txt

```
1  Max a <EPL lemme="abandonner_la_partie" CATEPL="V" CAT="V">abandonné</EPL> <EPL CAT="DET">la</EPL> <EPL CAT="N1">partie</EPL>.
2  il a <EPL lemme="abandonner_les_lieux" CATEPL="V" CAT="V">abandonné</EPL> <EPL CAT="DET">les</EPL> <EPL CAT="N1">lieux</EPL>.
3
4
```