Building a lexical bundle resource for CAT and MT

Natalia GRABAR* & Marie-Aude LEFER**

* STL UMR8163 CNRS, Université de Lille 3
** Marie Haps School of Translation and Interpreting, Brussels

Starting-point assumption

- General language bilingual lexical resources are mainly restricted to single words and compounds
  - Cf. Granger & Lefer (2012, 2013) on EN-FR bilingual dictionaries
- Terminological resources, though containing numerous MW terms, fail to include MWUs that are used to
  - express stance (i.e. attitudes and degrees of certainty, e.g. *it is very important that*, *it seems to me that*)
  - structure texts (e.g. *and that is why*, *when it comes to*)

Lexical bundles

"*recurrent expressions, regardless of their idiomaticity, and regardless of their structural status.* […] *sequences of word forms that commonly go together in natural discourse*"

(Biber et al. 1999: 90)

Functional taxonomy (Biber et al. 2004)

- 3 major discourse functions
  - Discourse organizers reflect relationships between prior and coming discourse
    - e.g. *and that is why, if you look at, on the other hand, when it comes to*

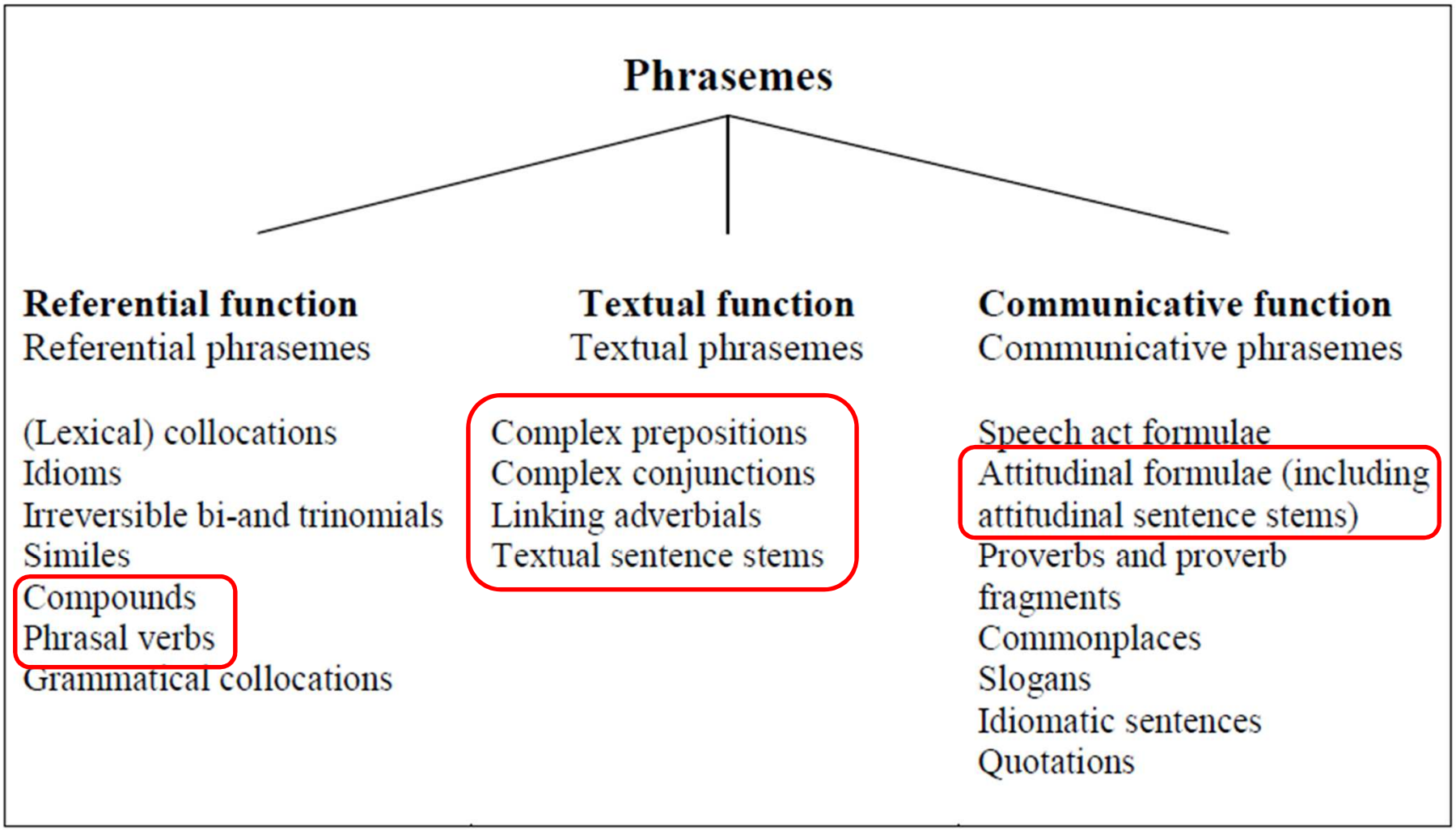
Functional taxonomy (Biber et al. 2004)

- 3 major discourse functions
  - Stance expressions express attitudes or assessments of certainty that frame some other proposition
    - e.g. *I don't know why, it is very important that, it seems to me that, you might want to*

Functional taxonomy (Biber et al. 2004)

- 3 major discourse functions
  - Referential expressions make direct reference to physical or abstract entities, or to the textual context itself
    - e.g. *those of you who, or something like that, a little bit more, at the same time, in the European Union, weapons of mass destruction*

The phraseological spectrum
(Granger & Paquot 2008)



n-gram extraction procedure

- Powerful discovery procedure that gives access to a whole range of recurrent sequences
- But manual filtering & deduplicating needed to keep structurally complete units of meaning only
  - E.g. *is why* > *this is why*, *that is why*, *which is why*; *for a long* > *for a long time*

Phraseology in CAT

- "*Whereas multi-word units are linguistically heterogeneous, in translation they raise a very similar set of problems. In order to translate them, they first have to be* ***recognized*** *as belonging together*" (Fernández Parra & ten Hacken 2008)
- "*formulaic expressions cannot be relied on to be translated compositionally but have to be considered* ***holistically***" (ten Hacken & Fernández Parra 2008: 3)
  - FR *ou encore*: EN **or even* (vs. *and*, *or*) (Granger & Lefer 2013)

Phraseology in CAT

- Use terminology tools in CAT software to extract MWUs and improve their representation/translation

Phraseology in Machine Translation

- "*In spite of the recent positive developments in translation technologies, multi-word units still present unexpected obstacles to Machine Translation and translation technologies in general, because of intrinsic ambiguities, structural and lexical asymmetries between languages, and cultural differences.* ***Multi-word unit identification and translation problems are far from being solved and there is still considerable room for improvement***" (Monti et al. 2013: 8)

Methodology

- Comparable & parallel corpus data
  - Cf. EN & FR are resource-rich languages

- NLP methods & manual validation
  - Build a lexical resource oriented towards CAT and post-editing of MT output
  - Amend or discard anomalous items from the automatically extracted bundle lists
    - "*Of course, it is one thing to rapidly create translation assets such as bilingual termbanks, and another entirely to ensure the quality of such resources*" (Haque et al. 2014: 46)

Methodology

- Step 1: automatic extraction of bundles in EN and FR
  - Comparable corpus of original texts, representing different genres: transcripts of EU parliamentary debates, research articles, news, editorials (ca. 7m tokens in total)

"Multi-word units belong to the general expressive means of a language. Although some of them are marked for register or text type, many are entirely unmarked. It is therefore not possible to collect a relatively small subset of multi-word units that are most likely to occur in a particular ST. No criteria comparable to the subject field for terminology can be used" (Fernández Parra & ten Hacken 2008)

Methodology

- Step 1: automatic extraction of bundles in EN & FR
  - Comparable corpus of original texts, representing different genres: transcripts of EU parliamentary debates, research articles, news, editorials (ca. 7m tokens in total)
  - Partial lemmatization in FR
  - New n-gram extraction method: 3-grams + longer n-grams containing them (= 'bundle families')
    - E.g. *on the other* > *on the other side*, ***on the other side of***, *on the other side of the*, ***on the other hand***, *on the other hand the, on the other hand there*
  - Analysis restricted to bundles that are found in at least 3 genres; low frequency thresholds

Methodology

- Step 2: manual selection of structurally complete bundles
- Step 3: automatic extraction of TL equivalents
  - Parallel corpora aligned at word level with Giza++ (Och & Ney 2000)
- Step 4: manual validation of TL equivalents

Comparable data extracted

	FRENCH	ENGLISH
Bundle families	**3251**	**1600**
Average size of bundle families	3.0 bundles/family	2.4 bundles/family
Largest bundle family	64 bundles	44 bundles
Selected bundles (after manual validation)	**1240**	**836**
Average length of selected bundles	3.6-gram	3.4-gram

Acquiring translation equivalents: a case study

- Monodirectional: EN to FR
- Corpus used: 'directional' Europarl (Cartoni & Meyer 2012)
- 400 EN discourse organizers and stance expressions + their FR equivalents
  - Analysis limited to equivalents with min. freq. = 2 (hapaxes were discarded)
  - 4000+ FR equivalents

DISCOURSE ORGANIZERS	ORIGINAL ENGLISH
Adding information	*there will also be, as well as, in addition to*
Comparing & contrasting	*in the same way, is not just about, on the other hand*
Summarizing & drawing conclusions	*at the end of the day, so it would be*
Exemplifying	*a good example of, among other things, areas such as, issues such as*
Expressing cause & effect	*one of the reasons why, this is not because, as a result, that is why*
Introducing topics & ideas	*the question of whether, when it comes to, the idea that, on the issue of*
Listing items	*the first is that, then there is, in the first place*
Paraphrasing & clarifying	*is not to say that, in other words, that does not mean*
Reporting & quoting	*he said that, in the words of, according to*

	ORIGINAL ENGLISH
STANCE EXPRESSIONS	*it is clear that, it is difficult to, it is necessary to, it is not surprising that, it is true that, it may well be, it would be wrong to, there is no doubt that, the truth is that, the problem is that*

Precision

	%
Discourse organizers	25.9
Stance expressions	32.3
Overall	27.7

Promising results

- Whole range of equivalents for many discourse organizers and stance expressions (1186 bundle pairs)
  - *among other things*: entre autres, notamment, entre autres choses
  - *but in the end*: mais finalement, mais en fin de compte, mais au final
  - *that is why*: c'est pourquoi, c'est la raison pour laquelle, voilà pourquoi, c'est pour cette raison, c'est pour cela, par conséquent
  - *it is clear that*: il est clair que, il est évident que, il ne fait aucun doute que, il apparaît clairement que, de toute évidence, il est manifeste que, il va sans dire que, à l'évidence, clairement
- Only 32/400 (8%) bundles with no FR equivalent

However…

	ORIGINAL FRENCH (4 genres)	TRANSLATED FRENCH (Europarl only)
Discourse organizers	438	135
Stance expressions	115	10
Other	687	12
TOTAL	1240	157

Stance expressions

- Found in Original FR & Translated FR: *c'est vrai que, de fait, en l'occurrence, il est clair que, il est évident que, il est possible que, il est vrai que, il ne faut pas, il n'est pas certain, penser que*

- No common bundles with *on* and *nous*

Stance expressions

- Typical structure of stance expressions: subject + V
  - Cross-linguistic contrasts: *it* = *il, elle, ce, cela, ceci, celui-ci, celle-ci,* etc.
- But
  - Better results for [subject+V+object/nominal predicate] stance expressions
    - E.g. *it will not be easy*: ce ne sera pas facile, il ne sera pas aisé; *it would be wrong to*: il serait erroné, il serait faux, il serait malvenu
  - Interesting (im)personal alternations
    - E.g. *it is hard to*/on a du mal à, *it is important to*/nous devons, *there is a need for*/nous avons besoin

Translation challenges for human translators, CAT & MT

- Polyfunctional bundles

- Categorial changes

Polyfunctionality of bundles

- *"It is not rare that a lexical bundle has more than a single function"* (Lee 2013: 380)
- Illustrations
  - *as far as*
    - Locative meaning → literal translation: *aussi loin*
    - Discourse organizer (topic introducer) → *en ce qui concerne, pour ce qui est de, s'agissant, concernant, quant à, pour ce qui concerne, au sujet de, en matière de*
  - *at the end of the day*
    - Temporal meaning → literal translation: *à la fin de la journée*
    - Discourse organizer →*au bout du compte, en fin de compte, finalement*

Categorial changes

- *away from*: éloigner, fuir, renoncer, abandonner
- *he said that*: selon
- *the first is that*: premièrement
- *is likely to*: probablement, vraisemblablement
- *there is no doubt that*: indubitablement
- *it is true that*: certes
- *we want to*: notre volonté

Next steps

- Test other NLP methods to identify target language equivalents
  - Using both parallel and comparable corpora
  - Relying on parallel corpora other than Europarl (use of TM)
  - Applying cleaning techniques to reduce noise (cf. Aker et al. 2014)
- Use the corpus data to build an EN><FR bilingual lexicon for CAT & MT

Thank you!

Merci !

References

Aker, A., Lestari Paramita, M., Pinnis, M. & Gaizauskas, R. (2014). Bilingual dictionaries for all EU languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 26-31.

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at* … Lexical Bundles in University Lectures and Textbooks. *Applied Linguistics* 25, 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Cartoni, B. & Meyer, T. (2012). Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *8th International Conference on Language Resources and Evaluation (LREC)*.

Fernández Parra, M. & ten Hacken, P. (2008). Beyond Terms: Multi-Word Units in MultiTerm Extract. *Translating and the Computer 30: Proceedings of the Thirtieth International Conference on Translating and the Computer*, 27-28 November 2008, London.

Granger, S. & Lefer, M.-A. (2013). Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries. In Aijmer, K. & Altenberg, B. (eds), *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson*. Amsterdam& Philadelphia: Benjamins, 157-176.

Granger, S. & Lefer, M.-A. (2012). Towards more and better phrasal entries in bilingual dictionaries. In Vatvedt Fjeld, R. & Torjusen, J.M. (eds), *Proceedings of the 15th EURALEX International Congress*. Oslo: UiO, 682-692.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In Granger, S. & Meunier, F. (eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: Benjamins, 27-49.

Hacken, P. ten & Fernández Parra, M. (2008). Terminology and Formulaic Language in Computer-Assisted Translation. *SKASE Journal of Translation and Interpretation* 3, 1-16.

Haque, R., Penkale, S. & Way, A. (2014). Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology*. 42-51.

Lee, C. (2013). Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives: Studies in Translatology* 21(3), 378-395.

Monti, J., Mitkov, R., Corpas Pastor, G. & Seretan, V. (eds) (2013). *Multi-Word Units in Machine Translation and Translation Technologies Workshop Proceedings*. Machine Translation Summit XIV. Allschwil: The European Association for Machine Translation.

Och, F.J. & Ney, H. (2000). Improved Statistical Alignment Models. In *Proceedings of ACL*. 440-447.