N-grams in multilingual corpora: extracting and analyzing lexical bundles in contrastive studies

Marie-Aude LEFER* & Natalia GRABAR**

* Marie Haps School of Translation and Interpreting, Brussels

** STL UMR8163 CNRS, Université de Lille 3

Outline

- Setting the scene
- Aims of the present study
- Data & method
- Discussion of the main results
- Concluding remarks & outlook

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
  - 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*

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*

n-gram extraction method

- Recurrent strings of *n* contiguous words, i.e. *n*-word sequences (2-grams, 3-grams, 4-grams, 5-grams, etc.)
- Frequency cut-off (e.g. 40 times per million words)

Lexical bundles in contrastive studies

- Combining the fields of **phraseology** and **corpus-based contrastive linguistics** is "*entering relatively unexplored territory*"

(Ebeling & Oksefjell Ebeling 2013: 1)

Lexical bundles in contrastive studies

- *"lexical bundles are a powerful window onto pragmatics and rhetoric. It is undeniably a quick-and-dirty method, but one that has great heuristic power: it generates a multitude of word sequences that have so far received very little interest in the contrastive literature"*

(Granger 2014: 69)

Lexical bundles in contrastive studies

- Granger (2014) on stems in EN & FR parliamentary debates and editorials
  - There are significantly more shared stems in FR parliamentary debates and editorials (56%) than in EN (29%)
  - Pervasiveness of multiword sequences in argumentative discourse in FR

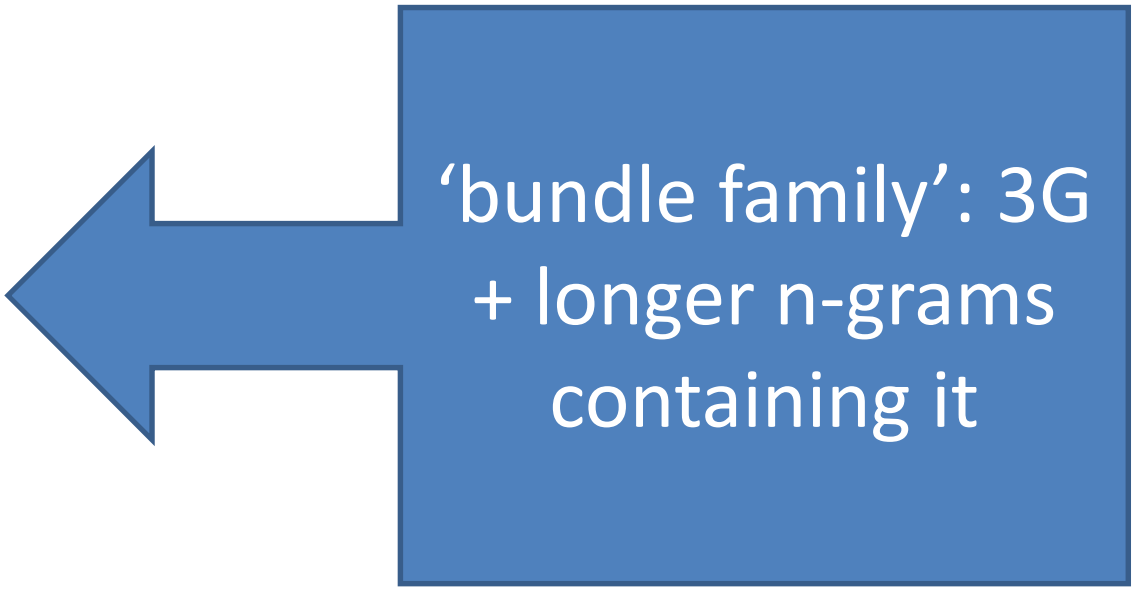
Cross-linguistic comparability in LB studies

- **Length of the bundles**
  - Equivalent bundles across languages can differ in length (Ebeling & Ebeling 2013, Granger 2014, Čermáková & Chlumská 2015)
    - E.g. *he said to himself – řekl si*; *for the first time – poprvé*
  - Limiting the analysis to one bundle length jeopardizes the cross-linguistic comparability of the data used

Aims of the present study

- (1) New n-gram extraction method for corpus-based contrastive studies
  - Reduce time-consuming manual weeding out, merging & deduplicating due to bundle incompleteness/overlap

at the end of
at the end of last
at the end of the
at the end of the day
before the end of
by the end of
by the end of the
by the end of the year
by the end of this
of the end of
since the end of
the end of
the end of a
the end of last
the end of last year
the end of the
the end of the day
the end of the year
the end of their
the end of this
the end of this year
towards the end of
until the end of



Examples of ‘bundle families’

- *of thousands of*, **tens of thousands of**, **hundreds of thousands of**

- ***as a result***, **as a result of**, as a result of the, as a result the

- *one way or*, **one way or the other**, **one way or another**

Examples of ‘bundle families’

- *d’une certaine*, **d’une certaine façon**, **d’une certaine manière**

- *l’expression de*, reprendre l’expression de, **pour reprendre l’expression de**, **selon l’expression de**, l’expression d’un, l’expression de l’

Aims of the present study

- (2) New extraction method tested in an illustrative EN-FR contrastive case study
  - Bundle pervasiveness across languages
    - Do written genres share more lexical bundles in FR than in EN? (Cf. Granger 2014)
    - Are bundles more pervasive in FR?

Overarching aims of the research

- Generate hypotheses for Corpus-Based Translation and Interpreting Studies
  - Source language interference *vs*. normalization
    (cf. Lefer & Vogeleer 2013)

- Build a corpus-informed lexical resource for CAT and MT (cf. MUMTTT workshop)

Corpora used

- Four comparable corpora
  - PLECI-News: news items from daily quality papers and magazines
  - Mult-Ed: newspaper editorials (opinion columns)
  - KIAP: research articles in medicine, economics and linguistics (Fløttum et al. 2006)
  - Europarl: transcripts of EU parliamentary debates (Koehn 2005, Cartoni & Meyer 2012)
- 900,000 tokens per genre in each language (ca. 7 m. tokens in total)
- Original texts only

Methodology adopted

- Combination of NLP extraction methods and manual validation & analysis
- Step 1: partial lemmatization of the FR corpora
  - Obvious cross-linguistic differences: *de/d'/du – of*; *le/la/l' – the; un/une – a; que/qu' – that*
- Step 2: automatic extraction of bundle families in each genre
  - Rather low frequency thresholds (min. 5 occurrences per genre for the 3G and min. 2 occurrences per genre for the accompanying, longer n-grams)
- Step 3: automatic selection of the bundles that are shared across genres (i.e. found in at least 3 of the 4 genres investigated)
- Step 4: manual validation and classification

Data extraction output

- *over the past*/204/multed/ep/pnews, that over the past/8/multed/ep/pnews, over the past five/15/multed/ep/pnews, over the past few/30/multed/ep/pnews, **over the past five years**/21/multed/ep/pnews, **over the past few years**/24/multed/ep/pnews, **over the past decade**/48/multed/ep/pnews, **over the past year**/46/multed/ep/pnews

REFERENTIAL EXPRESSIONS	FRENCH	ENGLISH
Compounds	*âge de départ à la retraite*	*freedom of speech*
Collocations	*problèmes liés à*	*young men and women*
Multi-word verbs	*avoir lieu*	*get rid of*
Organizations, countries & people	*la Cour Pénale Internationale*	*International Atomic Energy Agency*
Place	*à l'échelle mondiale*	*elsewhere in the world*
Time	*après la chute du mur de Berlin*	*over the next few days*
Quantity	*des centaines de milliers de*	*too much of*
Imprecision	*deux ou trois*	*some sort of*

DISCOURSE ORGANIZERS	FRENCH	ENGLISH
Adding information	*il y a aussi, ou encore*	*there will also be*
Comparing & contrasting	*à l'inverse, il en va de même pour*	*in the same way, is not just about*
Summarizing & drawing conclusions	*en fin de compte, en tout cas*	*so it would be*
Exemplifying	*par exemple, l'exemple de*	*a good example of*
Expressing cause & effect	*c'est pour cette raison que, c'est pourquoi*	*one of the reasons why, this is not because*
Introducing topics & ideas	*en ce qui concerne, s'agissant de*	*the question of whether, when it comes to*
Listing items	*en premier lieu, la première est que*	*the first is that, then there is*
Paraphrasing & clarifying	*en d'autres termes, pour ne pas dire*	*is not to say that*
Reporting & quoting	*comme l'a dit, pour reprendre l'expression de*	*he said that*

Stance expressions

FRENCH	ENGLISH
ce serait une erreur, il est évident que, on peut s'interroger sur, je pense que, tout le monde sait que	*it is not surprising that, it may well be, there is no doubt that, the truth is that*

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

Genre-shared bundles in the two languages

	FRENCH	ENGLISH
Referential expressions	**673**	424
Discourse organizers	**438**	288
Stance expressions	115	109
Polyfunctional bundles	14	15
TOTAL	**1240**	836

Discourse organizers across FR & EN

	FRENCH	ENGLISH
Adding information	14	10
Comparing & contrasting	**63**	41
Summarizing & drawing conclusions	5	1
Exemplifying	6	9
Expressing cause & effect	44	37
Introducing topics & ideas	**45**	29
Listing items	5	4
Paraphrasing & clarifying	9	4
Reporting & quoting	9	6

Comparing & contrasting in FR

à la différence de, à l'image de, à l'inverse, à l'inverse de, alors même que, alors que, au contraire de, au même titre que, autre chose que, bien que, ce n'est pas la première fois que, cela ne veut pas dire que, comme ce fut le cas, comme cela a été, comme celui/celle de, comme c'est le cas, comme dans le cas, comme l'ont fait, comme tous les autres, comme tout le monde, conformément, contrairement à, dans le cas contraire, dans le même ordre, dans le même temps, d'autre part, de la même façon, de la même manière, de la même manière que, de même que, d'une part, en revanche, et d'autre part, et non de, et pas seulement, il en est de même pour, il en va de même pour, il n'en demeure pas moins que, il n'en reste pas moins que, mais ce n'est pas, moins que, n'a rien à voir avec, n'a rien de, n'est pas le même, n'est pas non plus, n'est pas tant, non pas de, non seulement, non seulement de, or c'est, or c'est bien, or il y a, ou en tout cas, par rapport à, pas moins que, pas plus que, plutôt que, rien d'autre que, si ce n'est, similaire à celui de, supérieur à celui de, sur le même plan, tandis que

Introducing topics and ideas

à cet égard, à la question de, à l'idée de, à propos de, dans le cadre de, dans le contexte de, dans le débat sur, dans le domaine de, dans le secteur de, dans les domaines de, dans les secteurs de, dans un contexte de, du point de vue de, d'un point de vue, en ce qui concerne, en la matière, en matière de, en termes de, il ne s'agit pas de, il s'agissait de, il s'agit de, la question de, la question est de savoir si, la question se pose, le débat sur, le fait de, le fait que, les questions de, l'idée de, l'idée que, lorsqu'il s'agit de, pour ce qui concerne, pour ce qui est de, quand il s'agit de, quant à, qu'il s'agisse de, s'agissant de, s'il s'agit de, sur ce point, sur la question de, sur le dossier, sur le plan, sur le plan de , sur le plan économique, sur le plan politique

Referential expressions across FR & EN

	FRENCH	ENGLISH
Compounds	**205**	70
Multi-word verbs	65	**80**
Collocations	8	8
Organizations, countries & people	51	27
Place	**85**	51
Time	**89**	67
Quantity	82	86
Imprecision	7	6

N+prep+N compounds in FR

accès aux soins, accord de paix, agence de notation, aménagement du territoire, baisse des prix, cas de figure, chaine de télévision, commission d'enquête, conditions de travail, coût du travail, création d'emplois, dépenses de santé, droit d'asile, droit de vote, droits de douane, économie de marché, état d'urgence, gestion de la crise, groupe de travail, liberté d'expression, lutte contre la corruption, marché de l'emploi, marge de manœuvre, mise en œuvre, mise en concurrence, organisation du travail, pacte de stabilité, peine de mort, projet de loi, qualité de l'air, taux de croissance

N+prep+N compounds in FR

accès aux soins, accord de paix, agence de notation, amé… figur… conc… d'em… vote… d'urg… liber… marché de l'emploi, marge de manœuvre, mise en œuvre, mise en concurrence, organisation du travail, pacte de stabilité, peine de mort, projet de loi, qualité de l'air, taux de croissance

freedom of speech, point of view, quality of life, rule of law, standard of living, amount of money

human rights abuses, security council resolution

Concluding remarks

- Extraction method proposed contributes to the frequency-driven approach to contrastive phraseology
- Bundles are more pervasive in FR than in EN
  - Cf. FR tendency to rely heavily on rhetorical markers, to be explicitly emphatic and, more generally, to be more verbose than EN (Granger 2014; see e.g. Vinay & Darbelnet 1995)
- Important implications for a wide range of applied fields and Translation & Interpreting Studies

Next steps

- Explore the data (e.g. high-frequency bundles, genre-specific bundles)
- Improve extraction output (order in which the bundles are presented within a given family, deduplication) to speed up manual analysis
- Assess the impact of full lemmatization on the extraction procedure
- Develop a user-friendly web interface

Thank you!

Merci !

References

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).

Čermáková, A. & Chlumská, L. (2015). Comparing the language of children's literature. Paper presented at ICAME36, Trier University, May 2015.

Ebeling, J. & Oksefjell Ebeling, S. (2013). *Patterns in Contrast*. Amsterdam & Philadelphia: John Benjamins.)

Fløttum, K., Dahl, T. & Kinn, T. (2006). *Academic Voices – across languages and disciplines*. Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14(1), 58-72.

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

Hanks, P. (2013). *Lexical Analysis. Norms and Exploitations*. Cambridge/London: The MIT Press.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: *MT Summit X*, 79-86.

Lefer, M.-A. & Vogeleer, S. (eds). (2013). *Interference and normalization in genre-controlled multilingual corpora*. *Belgian Journal of Linguistics* 27.

Vinay, J.-P. & Darbelnet, J. (1995). [1958]. *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam & Philadelphia: John Benjamins.