UiO : University of Oslo

Signe Oksefjell Ebeling & Hilde Hasselgård

# Learners' and native speakers' use of recurrent word-combinations across disciplines

# Focus

- The use of recurrent word-combinations in texts produced by novice writers – both learners and native speakers – across disciplines.

- What types of n-grams – in terms of form and function – are salient in the two disciplines?

# Background

- The study of recurrent word-combinations, or n-grams, is rewarding “because they give insights into important aspects of the phraseology used by writers in different contexts” (Scott & Tribble 2006: 132).

- Although not all such combinations are of phraseological interest (cf. Altenberg 1998), they serve as a useful starting point for an investigation of how student writers apply them across disciplines.

- “Bundles occur and behave in dissimilar ways in different disciplinary environments.” (Hyland 2008: 20)

# Research questions

- What discourse functions do the recurrent word-combinations have?
- To what extent are the same patterns and functions used by learners and native speakers?
- Is L1 background or discipline more decisive for the use of recurrent word-combinations and their functions?

# Material

**Two corpora of novice academic writing:**

- The *Varieties of English for Specific Purposes dAtabase* learner corpus: Norwegian advanced learners of English (VESPA-NO)

- The *British Academic Written English* corpus: native speakers of British English (BAWE)

| | Linguistics | | Business | |
|---|---|---|---|---|
| | Texts | Words | Texts | Words |
| VESPA-NO (L2) | 239 | 267,855 | 70 | 47,335 |
| BAWE (L1, BrE) | 76 | 167,437 | 64 | 141,249 |

Mark-up of VESPA and BAWE to exclude e.g. footnotes, block quotes and headlines.

# N-gram extraction

- Inspired by Stubbs & Barth's (2003) study on recurrent phrases as text-type discriminators

- Extract the 100 most frequent 3- and 4-grams in each sub-corpus, using WordSmith Tools
- Focus on 3- and 4-grams
  - based on Altenberg's (1998) findings that the majority of recurrent word-combinations cluster as 2-, 3-, or 4-grams, some as 5-grams, and very few as 6-grams;
  - and on Stubbs & Barth's (2003) findings that three-word and four-word chains are better text-type discriminators than e.g. two-word or five-word chains.

# Functional classification
## adapted from Moon (1998: 217-218)

Differs from Moon (1998) in taking the organizational function out of the ideational, and instead operating with a textual-organizational function (more in line with Halliday (e.g. 2004))

| | | Function | Example |
|---|---|---|---|
| ideational — experiential | informational | • stating proposition, conveying information | of the brain |
| interpersonal | situational | • relating to extralinguistic context, responding to situation | as in tager flusberg |
| interpersonal | evaluative | • conveying speaker's evaluation and attitude | is likely to |
| interpersonal | modalizing | • conveying truth values, advice, requests, etc. | we can see |
| textual | organizational | • organizing text, signalling discourse structure | in this paper |

# Hypotheses

- The types of n-grams may differ between learners and native speakers (cf. Hyland 2008: 7 f, 20);
- The types of n-grams may differ across disciplines (cf. Hyland 2008: 20);
- Learners will be more visible authors in their texts, which may also show up in their recurrent word-combinations (cf. Paquot *et al.* 2013);
- Linguistics students will use more organizational n-grams than business students (cf. Hasselgård 2013).

# Comparing L1 groups: learners vs. native speakers

## Linguistics

| | BAWE-ling 3-grams | VESPA-ling 3-grams | Fisher's exact test | BAWE-ling 4-grams | VESPA-ling 4-grams | Fisher's exact test |
|---|---|---|---|---|---|---|
| Informational | 46 | 57 | P > 0.05 | 42 | 49 | P > 0.05 |
| Situational | 1 | 0 | P > 0.05 | 4 | 0 | P > 0.05 |
| Evaluative | 26 | 10 | P < 0.01 | 30 | 16 | P < 0.05 |
| Modalizing | 16 | 9 | P > 0.05 | 11 | 14 | P > 0.05 |
| Organizational | 11 | 24 | P < 0.05 | 13 | 21 | P > 0.05 |
| | 100 | 100 | | 100 | 100 | |

- native speakers: significantly more evaluative
- learners: (significantly) more organizational

# Comparing L1 groups: learners vs. native speakers

## Business

| | BAWE-bus 3-grams | VESPA-bus 3-grams | Fisher's exact test | BAWE-bus 4-grams | VESPA-bus 4-grams | Fisher's exact test |
|---|---|---|---|---|---|---|
| Informational | 64 | 73 | P > 0.05 | 65 | 80 | P < 0.05 |
| Situational | 0 | 0 | | 1 | 1 | |
| Evaluative | 9 | 7 | P > 0.05 | 12 | 4 | P > 0.05 |
| Modalizing | 9 | 1 | P < 0.05 | 2 | 4 | P > 0.05 |
| Organizational | 18 | 19 | P > 0.05 | 20 | 11 | P > 0.05 |
| | 100 | 100 | | 100 | 100 | |

- native speakers: (significantly) more modalizing (but small numbers)
- learners: (significantly) more informational

# Learners: Shared n-grams across the disciplines (full list)

| | 3-grams: | 4-grams |
|---|---|---|
| informational | | |
| situational | | |
| evaluative | | *it is important to* |
| modalizing | *we can see* | *I would like to*<br>*we can see that* |
| organizational | *in this essay*<br>*it comes to*<br>*on the other*<br>*the other hand*<br>*when it comes* | *at the same time*<br>*in this essay I*<br>*on the other hand*<br>*the other hand is*<br>*this essay I will*<br>*when it comes to* |

6% of 3-grams and 9 % of 4-grams are shared.

# Features that are typical of the Norwegian learners

- Function
  - Ideational and textual
    - Generally use more informational n-grams than the native speakers
    - Use slightly more organizational n-grams than the native speakers
- Form
  - n-grams with author presence (as hypothesized):
    - i will look at; in this paper i; i would like to; i will discuss; we can see that
  - other n-grams that are sentence stems or rhemes (Altenberg 1998)
    - the [first/second] text is; is an example of; decisions are made, the boss has more
  - overuse of some n-grams:
    - when it comes to

# Native speakers: Shared n-grams across the disciplines (frequencies and examples)

| | 3-grams: | 4-grams |
|---|---|---|
| informational | **18**<br>*a number of, it is a, part of the, such as the, that it is …* | **8**<br>*at the end of, in the form of, the nature of the…* |
| situational | **0** | **0** |
| evaluative | **6**<br>*as well as, due to the, is important to, the fact that, the importance of…* | **5**<br>*as well as the, it is clear that, it is important to, the fact that the …* |
| modalizing | **4**<br>*be able to, can be seen, it can be, need to be* | **1**<br>*to be able to* |
| organizational | **5**<br>*a result of, as a result, in terms of, in this case, one of the* | **3**<br>*a result of the, as a result of, on the other hand* |

33% of 3-grams and 17% of 4-grams are shared.

# Features that are typical of the native speakers

- Function
  - Ideational and interpersonal
    - Use less informational n-grams than the learners, but it is still the predominant function
    - Generally use more evaluative and modalizing n-grams than the Norwegian learners
- Form
  - Non-personal (self) projection (e.g. *it is clear that, it is argued that*)
  - Complex noun phrases (e.g. *the majority of the, the nature of the, as a result of*)
  - N-grams that reflect passive voice (e.g. *it can be seen*)

# Comparing disciplines: learners

| | VESPA-ling 3-grams | VESPA-bus 3-grams | Fisher's exact test | VESPA-ling 4-grams | VESPA-bus 4-grams | Fisher's exact test |
|---|---|---|---|---|---|---|
| Informational | 56 | 73 | P < 0.05 | 49 | 80 | p< 0.0001 |
| Situational | 0 | 0 | | 0 | 1 | |
| Evaluative | 10 | 7 | p>0.05 | 16 | 4 | p<0.01 |
| Modalizing | 9 | 1 | p < 0.05 | 14 | 4 | p < 0.05 |
| Organizational | 24 | 19 | p>0.05 | 21 | 11 | p>0.05 |
| | 100 | 100 | | 100 | 100 | |

Business: significantly more informational
Linguistics: significantly more evaluative/modalizing

# Comparing disciplines: native speakers

| | BAWE-ling 3-grams | BAWE-bus 3-grams | Fisher's exact test | BAWE-ling 4-grams | BAWE-bus 4-grams | Fisher's exact test |
|---|---|---|---|---|---|---|
| Informational | 46 | 64 | p<0.05 | 42 | 65 | p<0.01 |
| Situational | 1 | 0 | | 4 | 1 | p>0.05 |
| Evaluative | 26 | 9 | p< 0.01 | 30 | 12 | p< 0.01 |
| Modalizing | 16 | 9 | p>0.05 | 11 | 2 | p<0.05 |
| Organizational | 11 | 18 | p>0.05 | 13 | 20 | p>0.05 |
| | 100 | 100 | | 100 | 100 | |

Business: significantly more informational (as in VESPA), more organizational grams (unlike VESPA), but not significant.
Linguistics: Significantly more evaluative/modalizing (as in VESPA)

# Linguistics: Shared n-grams across the L1 groups (frequencies and examples)

| | 3-grams: | 4-grams |
|---|---|---|
| informational | **16**<br>*that there are, the number of, the use of, part of the…* | **7**<br>*and the use of, at the end of, by the use of…* |
| situational | **0** | **0** |
| evaluative | **7**<br>*in the same, meaning of the, the fact that…* | **7**<br>*as well as the, in the same way, it is important to…* |
| modalizing | **6**<br>*be found in, can also be, can be seen, can be used…* | **5**<br>*can be found in, can be seen in, it is possible to…* |
| organizational | **7**<br>*an example of, in this case, in this essay, looking at the…* | **6**<br>*an example of this, example of this is, in this case the…* |

36% of 3-grams and 25% of 4-grams are shared.

# Features that are typical of linguistics

- Function
  - Ideational and interpersonal
    - Predominantly informational (most overlap between the L1 user groups)
    - Topic-specific
    - Generally more evaluative and modalizing n-grams than the business students
- Form
  - Complex noun phrases (e.g. *at the end of, by the use of, in the case of*)
  - N-grams with *can* predominate in the modalizing function (*can be found in, can be seen in, can also be*)

# Business: Shared n-grams across the L1 groups (full list)

| | 3-grams: | 4-grams |
|---|---|---|
| informational | *a lot of*<br>*that they are* | |
| situational | | *at the same time* |
| evaluative | *is important to*<br>*it is important*<br>*the importance of* | *it is important to* |
| modalizing | | |
| organizational | *based on the*<br>*in order to*<br>*there is a*<br>*one of the* | *on the other hand* |

9% of 3-grams and 3% of 4-grams are shared.

# Features that are typical of business

- Function
  - Ideational
    - Highly informational and topic-specific, even more so than was the case for linguistics
    - Some organizational features and very few of the others
- Form
  - Hardly any overlapping n-grams across the two L1 groups, apart from evaluative grams including *importan** (*is important to, it is important, the importance of*)

# Summary of findings: functional types of n-grams

- The n-gram approach and the classificatory framework enabled us to identify differences between both disciplines and L1 groups.
- The ideational/informational grams are typical for both L1 groups and both disciplines – above 50% in all subcorpora except NS linguistics
  - Not surprising, since academic disciplines have been found to be highly informational and we are dealing with novice academic writers.
- Situational n-grams were rare in all the corpora (found only in NS linguistics)
- Evaluative and modalizing n-grams were more frequent in linguistics than in business in both L1 groups
- Organizational n-grams were more frequent in linguistics in VESPA and more frequent in business in BAWE

# Summary of findings: Learners vs. native speakers

- Far fewer overlapping n-grams across the disciplines among the learners than among the native speakers.
- More overlapping n-grams between learners and native speakers in linguistics than in business.
  → The linguistics papers are more similar across the L1 groups than the business papers
- Some features of the distribution of *functional* types of n-grams mark learners off from native speakers in both linguistics and business:
  - The learners in both disciplines have fewer modalizing and evaluative grams.
  - A slight tendency for the learners to use more informational n-grams (significant only with 4-grams in business).
  - In linguistics, the learners have more organizational grams than native speakers, but in business they have slightly fewer.

# Summary of findings: Learners vs. native speakers (cont.)

- Differences in the *form* of the n-grams
  - Learners: more n-grams involving 1st person pronoun
  - Native speakers: more n-grams suggesting complex noun phrases and verb phrases with passive voice
  - Native speakers: more n-grams with non-personal projection (extraposition)

# Summary of findings: Disciplinary differences, linguistics vs. business

- There are more overlapping n-grams between the corpora in linguistics than in business.
- Linguistics students have fewer informational n-grams than business students (across L1 backgrounds)
- Linguistics students have more evaluative and modalizing n-grams than business students.
- The discipline comparison involved more statistically significant differences than the NS/NNS comparison.

# Further work

- More material from the business discipline (esp. learners)
- Comparison with other disciplines
- Comparison with other L1 backgrounds
- Comparison with ‘specialist’ writing in the same disciplines.

# Applications

- The development of a “multi-word academic word list”
- Disciplinary and L1-specific use of n-grams could feed into EAP courses & teaching materials.
  - Functional types of n-grams that differ greatly across L1 background (e.g. modal / evaluative; clausal vs. nominal n-grams)
  - Functional (and structural) types of n-grams that are typical for academic disciplines
- Findings indicate a greater need for more explicit instruction among the NNS business students.

## Works consulted


Ädel, A. & U. Römer (2012) Research on advanced student writing across disciplines and levels. Introducing the *Michigan Corpus of Upper-level Student Papers*. *International Journal of Corpus Linguistics* 17:1, 3-34.

Altenberg, B. (1998) On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press. 101-122.

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

Ebeling, S.O. (2011) Recurrent word-combinations in English student essays. *Nordic Journal of English Studies*, 10:1, 49-76.

Halliday, M.A.K. (2004) *An Introduction to Functional Grammar*. 3rd ed., revised by C.M.I.M. Matthiessen. London: Arnold.

Hasselgård, H. (2012) *Facts, ideas, questions, problems*, and *issues* in advanced learners' English. *Nordic Journal of English Studies*, 11:1, 22-54.

Hasselgård, H. (2013) Metadiscourse in novice academic English. Paper given at ICAME34, Santiago de Compostela.

Hyland, K. (2008). As can be seen. Lexical bundles and disciplinary variation. *English for Specific Purposes 27*, 4-21.

Moon, R. (1998) *Fixed Expressions and Idioms in English. A Corpus-based Approach*. Oxford: Clarendon Press.

Paquot, M., H. Hasselgård & S.O. Ebeling. (2013) Writer/reader visibility in learner writing across genres. A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In: S. Granger, G. Gilquin & F. Meunier (eds), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 377-387.

Scott, M. and C. Tribble (2006) *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam / Philadelphia: John Benjamins.

Stubbs, M. & I. Barth (2003) Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language*, 10 (1), 61-104.

# Corpora

BAWE, see http://wwwm.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx

VESPA, see http://www.uclouvain.be/en-cecl-vespa.html