Quantifying Lexico-Semantic Complexity in Academic Writing with Complexana

Christoph Haase
Purkinye University, Ústí nad Labem
Czech Republic

Abstract

In this contribution, a novel tool for automatically assigning a score for semantic complexity to a given text is introduced together with a number of results from corpus studies concerning the linguistic parameters of academic texts. The tool is based on the WordNet project as part of the semantic-web initiative. The texts come from a self-compiled corpus project called SPACE and represent different registers of academic writing – specified academic texts and popular-science articles. The comparability is ensured by selecting texts in parallel which concern the same academic research. Further, by application of the introduced tool, ideas are developed how a self-assessment of semantic complexity can be applied to teaching academic writing.

Key words: Knight errant, humor literature, romance chivalry, windmills, disappointment, adventure

Introduction

In this contribution, principles of corpus linguistics, in particular lexicostatistics methods will be applied to the study of English for Academic Purposes (EAP). In particular, the development of a software tool, Complex Ana, for Complexity Analyzer, will be the focus of attention as it connects a number of linguistic but also pedagogical approaches to academic writing. The motivation for the development of the tool grew out of observations of nonnative speakers (NNS) in their struggle with the norms and demands of academic text production. As modern corpora have been used by educators for some time in the teaching of academic writing, one goal was to help students get a better grasp and reflexivity of their own writing by using corpora not simply to query collocations and their use but also to compare the own writing against published sources and calibrate their revisions accordingly. The experience of the user is limited by the number of examples he or she can overlook and in this sense corpora can help educators and learners to extend their view by giving authentic examples. This encourages principles of inductive learning on collocations that appear with frequencies higher than chance and in opposition to deductive, rule-based learning, this type of learning enables learners to compose acceptable and publishable results (cf. Haase, 2015).

In this contribution we will investigate which attempts have been made in order to automatically profile texts according to register and genre according to lexicostatistic tools and what the component of corpus linguistics in this can be. Further, we will present data from a survey of texts in a parallel corpus called SPACE that will show that reliable results in profiling texts can be obtained when a defined variable – lexicosemantic complexity – is considered as the decisive criterion of text differentiation. The corpus will be introduced and the development and application of a novel tool called ComplexAna in its impact on the study of academic language and education will be discussed.

2. Corpora and the genre discussion

A shift in the study of English for Academic Purposes was precipitated by the approach by Swales who defined genre and the knowledge of genre as an integral part of the competence of an author to produce adequate texts (Swales 1990). Until then, the approach prevailed that ‘good English’, preferably that of a native speaker (NS), was considered sufficient for academic publication practice and this is reflected in the fact that numerous prestigious journals would keep NS editorial staff and a sort of linguistic ‘gate keeping’ took place that put NNS at a disadvantage. The genre approach led to a reappraisal of linguistic-pragmatic skills that also considered the knowledge of the conventions as valuable and in fact almost as important as the ability to formulate coherently in the second language English.

The conventions are to be found in parameters like 'setting', 'style', 'variety', 'text type' and 'domain' (Baker 2010: 44). Out of this approach, which was on the grammatical-linguistic side informed by Biber's attempts (2006) at register differentiation and on the socio-linguistic side by the English as an international linua franca movement (ELF), which considers NNS Englishes as equivalent varieties and which tries to overcome the NS gate keeping, especially in academia, the understanding grew that academic texts are characterized by a larger number of criteria than originally assumed. This is reflected in the rise of corpus-linguistic methods and the use of corpora in teaching. Connor summarizes these demands in the following points

1. Growing recognition that tasks and materials must be grounded in the analysis of real texts

2. Recent discourse analysis tools that allow us to see how discourse features are linked ''to issues of writer purpose, identity audience expectations, cultural schemata, disciplinary perceptions, and so on''

3. Growing curiosity about the ways EAP students actually write, and

4. Available technology, through corpus analysis of language use and patterns in student writing, for the purpose of studying systematic variation. (Connor 2004: 296).

In teaching academic writing, corpora are a relatively recent phenomenon and a large number of publications introduced new corpora for teaching, for example Aston, Bernardini & Stewart 2001, Braun, Kohn & Mukherjee (eds.) 2005, Granger 2004, Nesselhauf 2005, Renouf (ed.) 2006, Sinclair 2004, or the ReCall 19, 3 (2007) special issue on "Incorporating Corpora in Language Learning and Teaching". Thus, on the pedagogical side, an interest arose to apply corpus results to teaching, especially for English for Special Purposes (ESP) where corpora provide authentic teaching materials: "Corpus-based studies are especially amenable to the teaching of the reading and writing skills and the development of academic literacy" (Yoon & Hirvela 2004:258). The aforementioned highly specific and advanced forms of academic communication are addressed by the project described here which tries to connect corpora as example-generators towards corpora as an effortless teaching aid for learners and practitioners, NS and NNS alike. In the following, we introduce the SPACE corpus, an acronym for Corpus of "Scientific and Popular ACademic English", (published in Haase 2009, 2010, 2011 and 2013).

3. The SPACE corpus

3.1 Design

The first versions of the SPACE corpus were compiled in 2005 at Chemnitz University of Technology where it was also hosted back then. The original idea for the corpus was to offer a comprehensive view on the major disciplines of natural sciences ranging from the most abstract (quantum theory, particle physics and cosmology) to the more concrete (zoology, plant science). It would also cover the 'hard' and the 'soft' natural sciences for the maximally possible range of abstraction. An observed need at institutions of higher education for high-quality writing courses that could encompass academic writing was addressed with the inclusion of the corpus in teaching. As a result, students acquired skills from the practical work with the corpus that could be taught by general educators because the scientific disciplines are too diverse. A corpus that would offer a span of very different branches within the natural science therefore was a natural starting point but in view of the different text registers that find their place under the overall umbrella term 'academic writing', further components were added that made the corpus a true parallel and helpful teaching aid. It has been described in different publications (the standard paper on SPACE is Haase 2009) in greater detail therefore the following discussion is kept concise.

2.2 Corpus structure

In the study of English for Academic Purposes (EAP) it has become an important aspect to investigate text-type differentiation so that practitioners can orient on a standardized way of presenting research results, either by using the correct markers of stance or personal involvement, by using conventionalized lexical bundles or collocations that are within the expectations of the readership and even the way the authors themselves are represented in their text is subject to quantification and formalization. In addition to this, different 'house styles' of journals exist according to which authors need to orient. Taken all these aspects together, a relatively clear picture emerges of what characterizes the text type itself and it seems clear that no corpus can represent all types. The SPACE corpus therefore always followed a different approach. Given that two disciplinary branches existed, physics and biosciences, the attempt was made to complement these branches by comparable texts on a different level of professionality albeit still academic. This text type was found in popular-science texts that have a scientific counterpart in specialized journals.

While these original texts come from pre-print servers like arxiv and from openly accessible research results published in the Proceedings of the National Academy of Sciences (PNAS), the second tier of the SPACE corpus is comprised of texts that represent derivations from the first, the specialized tier. These derivations are regularly published in multi-disciplinary science journals like the New Scientist but in order to address a general-interest academic readership, these texts represent simplified summaries and journalistic interpretations of the original research results.

The structure of SPACE can be seen in table 1:

Table 1: Structure of the SPACE corpus: Subcorpora and sizes

Subcorpus	Descriptors	word count
arXiv	physics, astrophysics, quantum mechanics	809,320
New Scientist – physics	physics, astrophysics, computer science, quantum mechanics	203,470
Proceedings of the National Academy of Science (PNAS)	biochemistry, genetics, genetic engineering, microbiology	267,105
New Scientist - biosciences	biochemistry, genetics, genetic engineering, microbiology	30,499
Public Library of Science – Medicine (PLoS),	medicine, virology, clinical psychology, public health	217,254
New Scientist – medicine	medicine, virology, clinical psychology, public health	17,050
Total 1,544,149		

3.3 Corpus application in teaching

The character of the SPACE corpus as a teaching aid was enhanced by the inclusion of the popular-science texts as in the difference between the original and the journalistic interpretation, a large number of linguistic parameters can be assessed. When the linguistic foundations of argumentation at different levels of abstraction can be made transparent, the students can make informed decisions about which principles to use.

The following example shows at first a specialized segment, followed by its popularized counterpart:

0014AX experimentally it is hard to maintain two parallel plates uniformly separated by distances less than a micron. So one of the plates is replaced by a metal sphere of radius R where R » z. Here a sphere of radius R = 100 μm imprinted by sinusoidal corrugations was used instead of one of the plates (see Fig. 1). As both z « R and « R, the normal Casimir force can be calculated by use of the Proximity Force Theorem [29] as Fnor(z, ') = 2REcor pp (z, '). The accuracy of this theorem for our parameters is of order 0.2% (note that the recent result [30] claiming a worse accuracy for the PFT is applicable only to the pure nonrelativistic regime with separations z no larger than several nanometers and also small R).

0014NS Last week, Mohideen and his team announced that they had measured this lateral force. They placed two corrugated gold plates a few hundred nanometres apart with their peaks and troughs aligned (see Diagram). When they moved the plates slightly out of alignment, they detected a force of a few piconewtons that pushed them back into position.

As can be seen, the original text shows lexical specialization (or: lexico-semantic complexity), nominalizations and passive voice. The argumentation relies on cause-effect relationships. The popular text (0014NS) uses lexical items understandable for nonexperts, is written in active voice and the argumentation relies on a sequential narration of events. Thus, a wide spectrum of linguistic features can be studied in parallel comparison which is helpful for learners. In summary, the differences are (cf. Haase 2013):

a) Markers of propensity (the commitment of the author(s) to the validity of their results and findings (examples: modal verbs, modal auxiliaries, hedge expressions, see Haase, 2011 in Schmied (ed.) 2011)

b) Stylistic devices like amplifiers (examples: completely, absolutely) and boosters (examples: very, highly, immensely)

c) High lexical specialization (cf. table 2)

d) Linguistic markers of causality (causatives, resultatives, conjunctions, use of tense)

The lexical specialization which is the basis for the stud described here is illustrated in two parallel examples in table 2:

Table 2: Two parallel titles from SPACE: academic (PNAS) and popular-academic (NS)

Code source words title

0014AX arxiv 3236 Demonstration of the Lateral Casimir Force

0014NS

New Scientist 412 Out of the void

Table 2: Two parallel titles in their academic (PNAS) and popular-academic (NS) version

The academic title uses specialized terminology; the popular-science title is sensational. Teaching the ability to differentiate facilitates learners' access the reception of specialized academic texts and helps improve their skills in cohering to the requirements of the academic text type or genre. Further, it initiated attempts to formalize what is done by the science journalists in order to cross the barrier of complexity. The ComplexAna tool was developed to meet these demands in a transparent but simple way.

4. The lexico-semantic complexity analyzer

4.1 Development

In order to capture the systematicity of the process that is taken by the human science journalist when he or she writes a popularized summary of a highly specialized academic text we found it helpful to define a parameter of so-called lexico-semantic complexity. Words for highly specific objects and events exist on a scale of lexical specialization and a common semantic core in it would have a lower level of complexity than a specialized term. This core may be located at the category of base-level (cf. Evans & Green 2006:248). This can be illustrated with the terms taken from two parallel texts (corpus codes 0007AX and 0007NS) in table 3. The table also shows markers of vagueness.

Table 3: Different lexico-semantic complexity in two parallel examples from physics

	Academic text 0007AX	**popular academic text0007NS**
markers of specialization	conjectures, compactification, coalescence, planetesimals, angular, mesoscopic, gauge field, accretion, radial drag	dead stars, cloud of gas, hot star, proto-planetary disc, rogue comets
markers of vagueness	suggest X may have, should detect Rc, deviations are weak, may be turbulent it	may be hard, can be slow, they probably rebound, could charge up

If we therefore assume that the complexity can be seen as a hint of the argumentative strength in an academic text then we can use this to systematize this as a lexico-semantic function and use it in an automatic profiling text for learners. It could help compare texts and measure their experienced difficulty and also generate learner data from recognition tests to correlate them with words which are 'felt' to be difficult. In a further step, a text can then be re-formulated by the learner who can measure if his or her text meets the demands of academic writing by using higher-specialized vocabulary. The transfer may be accompanied in both directions, upward to higher specialization (higher lexico-semantic complexity) and downward to lower specialization (lower lexico-semantic complexity).

We therefore define lexico-semantic complexity as an dimensionless aggregated score combining lexical specificity with a number of syntactic values like sentence length given by the number of (sub)clauses and lexico-statistic values like the number of words in a text that is unknown to the WordNet ontology (described below). In consequence, the scores of semantic complexity can be calculated. They make sense in comparison only as they are relative, not absolute scores. The mostz important component for the complexity score of course needed to be a variable that would have an external support by comparison with lexical material that could be considered common core. A number of variables was tested and rejected, for example the comparison with the academic wordlist (Coxhead 2000), frequency of the items and their keyness. An extensive and well-researched data basis was found in the linguistic ontology of WordNet (www.wordnet.org) which allows implementation in different tools and has the backing of a longterm project at Princeton University. An example entry from WordNet looks as follows:

Figure 1: Display of a queried term in WordNet

The Word Net project was started in the late 1980s and the first version Word Net was published in 1991. It contains compounds, phrasal verbs, collocations, and idiomatic phrases; the word is the basic unit. Word Net does not decompose words into smaller meaningful units a, though a comparison with componential analyses reveals some common aspects (Fellbaum 1998: 3).

For the automatic generation of a complexity score, the ontological information was most crucial as the ontology parses nouns into their hyponyms and hyperonyms which create a tree-like structure. The position of a lexical item in this structure can be used to gauge the ontological position of a lexical item. We developed a software tool that could read out the position of any queried lexical item in this structure originating from Word Net.

The Complexity Analyzer (Complex Ana) was written in Perl as a standalone application.

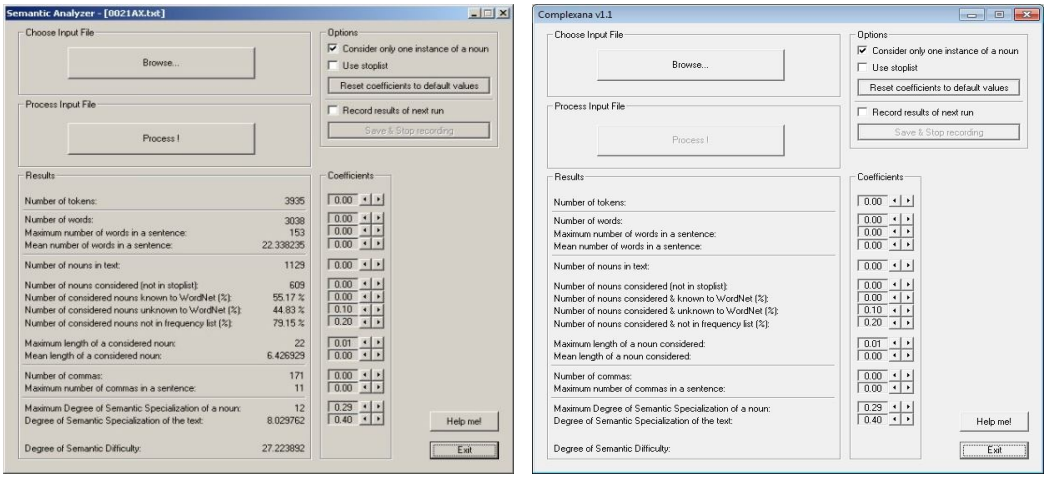

Figure 2: Initial and current version of ComplexAna

Several computational tasks are carried out by the tool if a text is run. Learners have been using its functionality to test their own texts and modify them accordingly to adapt to standard scores of complexity established over the runtime of the project.

4.2 Complex Ana algorithm and data discussion

The tool provides a user interface that enables a free adjustment of all parameters according to which should be weighted stronger or weaker. As it was written in Perl, a license-free implementation of Word Net could be used, in fact the use is encouraged by the WordNet creators. We managed to create the ComplexAna tool as a self-contained executable file that needs as its only requirement a Perl installation on the home directory.

In the first step, the tool requires the input of a raw text (txt format). The text is read by the tool and part-of-speech (POS) tagged using TreeTagger which provides overall robust accuracy of ca. 96% and it is part of the package as well. A side effect is that all tagged files can be retrieved by the user and thus ComplexAna is also a simple POS-tagger. This is shown for a segment from the discussed corpus file 0014AX:

The	DT	the
measured	VVN	measured
force	NN	force
shows	VVZ	show
the	DT	the
required	VVN	require
periodicity	NN	periodicity
corresponding	JJ	corresponding
to	TO	to
the	DT	the
corrugations	NNS	corrugation
.	SENT	.

At the same time, types and tokens in the texts are quantified and recorded. The tagging is necessary because the score of lexico-semantic complexity is calculated only for the nouns in the text. There is no consideration of semantically complex verbs. All nouns are written into a separate file which can also be retrieved. We further added functionality for stoplists in order to ignore items that generate false scores or in any way influence the overall statistics negatively. In the next step, all nouns from the text are cross-checked with the implemented WordNet ontology. The aggregated positions of the nouns in WordNet is used to calculate to most important influencing parameter. This score is subjected to a number of correctional terms which can be seen in figure 3 below:

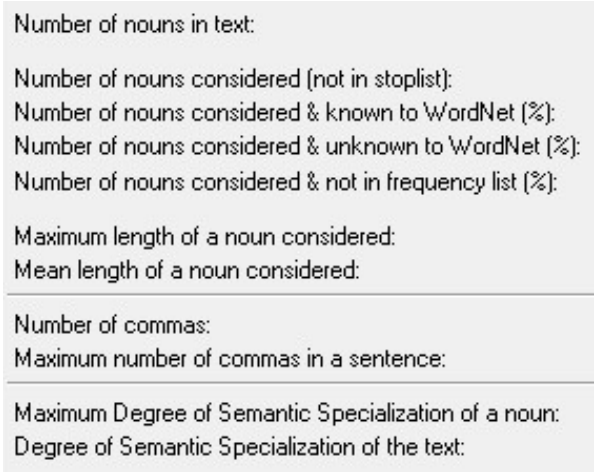

Figure 3: Parameters influencing the lexico-semantric score in ComplexAna

The parameters are used for correction terms that influence the main parameter. As a last step, a score is generated that can stand for the overall lexico-semantic complexity of the text, a dimensionless number. This is done exemplarily in the data given below:

Table 4: ComplexAna scores for academic (AX &PN) and popular-science texts (NS)

DESCRIPTION	0009AX	0009NS	0010PN	0010NS	0013AX	0013NS	0037PN	0037NS.txt
Tokens	6329	301	2431	288	2057	384	5857	272
Words	4861	251	2061	241	1569	345	4383	213
words/sentence (max)	112	32	70	31	82	37	78	28
words/sentence (mean)	17.1162	20.916667	24.535714	20.083333	16.010204	23	19.30837	21.3
nouns/text	1947	93	648	77	589	105	1820	88
nouns analyzed	583	69	265	65	243	64	661	74
nouns in WordNet (%)	80.96	72.46	86.04	81.54	70.78	85.94	73.07	82.43
nouns unknown (%)	19.04	27.54	13.96	18.46	29.22	14.06	26.93	17.57
nouns not in freq list (%)	58.66	66.67	49.81	64.62	67.9	65.63	66.41	60.81
Noun length (max)	16	14	24	15	19	11	33	23
Noun length (mean)	7.253859	6.072464	7.626415	6.4	5.839506	6.234375	8.003026	6.837838
Commas	349	11	112	17	88	9	476	15
commas/sentence (max)	7	3	13	3	13	2	30	4
Semantic Specialization of a noun	15	11	13	14	14	13	14	14
Degree of Semantic Specialization of the text	8.141949	7.66	8.02193	8.716981	8.05814	8.127273	8.221532	8.344262
Degree of Semantic Difficulty	**22.40314**	**21.480957**	**22.577262**	**18.466023**	**23.975313**	**21.662159**	**23.654407**	**21.546624**

As can be seen, in last row the differing scores of overall lexico-semantic complexity show contiguity but there are systematic differences. This shows that the score in itself needs to be complemented with other parameters as the texts differ in length and a short text of medium specialization would score higher than a longer text with (impressionistically) higher specialization – simply because it contains more low-level 'filler' nouns. Added parameters are for example the number of subclauses and sentence length measured in words between the stop signals. This explains why the popular-science text 0009NS with less words included in WordNet obtains a lower overall score than its academic counterpart 0009AX. The addition of these parameters achieved a score balancing that meets the impressions of the students.

Compared for the entire SPACE corpus, the results are strikingly systematic:

Table 5: Complex Ana scores for all subcorpora in SPACE

Type/Domain	**Physics/Astrophysics**	**Biosciences**	**Psychology**
Specialized academic	23.61	26.28	22.37
Popular science	19.11	19.79	19.36

The survey shows also that even though the popular-science texts vary in length, their scores are on average below 20 (19.11 for the popular physical texts 0001NS – 0046NS and 19.79 for the popular bio/genetic texts 0047NS – 0107NS). In a direct domain comparison biosciences have the highest level of lexico-semantic complexity (26.28), psychology as a social science has been included here for comparison, is considerably lower (22.37) while physics seems closer to psychology in complexity than to the biosciences.

5. Conclusion

Integrating the use of corpora and the findings from corpus data in teaching has beneficial effects for novice practitioners in EAP but in this contribution the extended suggestion is to also integrate corpus tools like ComplexAna for learners to fine-tune their own text, thus improving readability and style and comparing them to the texts o f published authors. Therefore, corpora can be more than simply repositories of best-example practice. A corpus like SPACE and an adjoining tool like Complex Ana can helps students and learners to optimize their academic competence but also for the researcher to produce texts meeting the requirements of the genre. To be able to calibrate the lexico-semantic complexity of their own writing and of the writing of others is on the one hand side a learning aid that has been used in practice. On the other hand, linguists can use the tool to prove register differences in-between text types and thus solidify anecdotal evidence on a given text by falsifiable data.

References

Aston, G., Bernardini, S., and Stewart, D. (2001). Corpora and language learners. Amsterdam: John Benjamins.

Aston, G. (2002). The learner as corpus designer. In: Kettemann, B., and Marko, G. (eds.), Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 9-25.

Baker, P. (2010). Sociolinguistics and corpus linguistics. Edinburgh: Edinburgh University Press.

Biber, D. (2006). Stance in spoken and written university registers. Journal of English for Academic Purposes 5, 97–116.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents, ReCALL 17 (1), 47-64.

Braun, S., Kohn, K., and Mukherjee, J. (2006). Corpus technology and language pedagogy. Frankfurt am Main: Peter Lang.

Chambers, A. (2005). Integrating corpus consultation in language studies. Language Learning & Technology 9 (2), 111-125.

Connor, U. (2004). Intercultural rhetoric research: beyond texts. Journal of English for Academic Purposes 3, 291–304.

Coxhead, A. (2000). A new academic word list. TESOL Quarterly 34, 213–238.

Evans, V. and Green, M. (2006). Cognitive linguistics. An introduction. Edinburgh: Edinburgh University Press.

Fellbaum, C. (1998). Wordnet. An electronic lexical database. Cambridge: MIT Press.

Granger, S. (2002). A bird's eye view of learner corpus research. In: Granger, S., Hung, J., and Petch-Tyson, S. (eds.), Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins, 3-33.

Granger, S. (2004). Computer learner corpus research: Current state and future prospects. In: Connor, U., and Upton, T. (eds.), Applied corpus linguistics: A multidimensional perspective. Amsterdam: Rodopi, 123-145.

Granger, S., Hung, J., and Petch-Tyson, S. (eds.) (2002). Computer learner corpora, second language acquisition and foreign language teaching. Amsterdam: John Benjamins.

Haase, C. (2006). A Crosslinguistic View on Causativity: Causer Neglect. R. Povolná and O. Dontcheva Navratilová (eds.) Discourse and Interaction 2. Brno Seminar on Linguistic Studies in English: Proceedings. Brno: Masaryk University, 57-70.

Haase, C. (2008). Hedging in Academic Learner English: A Survey of German and Czech Students. In: Tollet, A. (ed.) English Studies & Language Teaching. Plzen: University of West Bohemia.

Haase, C. & Schmied, J. (2008). Clause linking in specialized and popular academic English: An investigation into scientific texts from the space corpus. In: Lewandowska-Tomaszczyk, B. (Ed.): Corpus Linguistics, Computer Tools, and Applications - State of the Art. Frankfurt/New York: Peter Lang, 2008.

Haase, C. & Schmied, J. (2008). English projects in teaching and research in Central Europe. Cuvillier, Göttingen.

Haase, C. (2008). Linguistic Determinants of English for Academic Purpose. In: Frankenberg-Garcia, Ana et al.: Proceedings of the 8th teaching and language corpora conference. - Lisbon : Associaocao de Estudos de Investigacao Cientifica, 2008, S. 128 – 132.

Haase, C. (2009). Pragmatics through corpora in cultures: An empirical comparison of academic writing. Topics in linguistics 4, 23-30.

Haase, C. (2010). Mediating between the "two cultures" in academia: The role of conceptual metaphor. Discourse and Interaction 1 (3) (2010), 5-18.

Haase, C. (2011). Second language learner processing of hedge expressions. Evidence from Academic English. In C. Haase & N. Orlova (Eds.) (2011). ELT – Converging approaches and Challenges. Newcastle: Cambridge Scholars, 41-56.

Haase, C. (2011). Modal Indeterminacy and Evidentiality in Adverbial Expressions: A Culturome in Academic Writing? In J. Schmied (Ed.). Academic Writing in Europe: Empirical Perspectives. Göttingen: Cuvillier, 51-64.

Haase, C & Schmied, J. (2011). Conceptualising Spatial Relationships in Academic Discourse: a Corpus-Cognitive Account of Locative-Spatial and Abstract-Spatial Prepositions. In Explorations across Languages and Corpora, ed. by Gozdz-Roszkowski, Stanislaw. New York, Frankfurt: Peter Lang, 351-370.

Haase, C. (2013). Tools for identifying and teaching semantic complexity in academic writing In: Language & Technology: Computer Assisted Language Teaching, edited by Darah Tafazoli. Tehran: Khate Sefid, pp.129-137.

Haase, C. (2015). Strata of academic English in corpora – A parallel genre approach. In Monika Černá, Jaroslava Ivanová & Šárka Ježková (Eds.). Learner Corpora and English Acquisition. Pardubice: University of Pardubice Press, 81-90.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge: Cambridge University Press.

Hyland, K. (1994). Hedging in Academic Writing and EAP Textbooks, English for Specific Purposes 13 (3), 239-256.

Kaltenböck, G., and Mehlhauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching, ReCALL 17 (1), 65-84.

McEnery, T., and Wilson, A. (1997). Teaching and language corpora, ReCALL 9 (1), 5-14.

Nesselhauf, N. (2005) .Collocations in a learner corpus. Amsterdam: John Benjamins.

Renouf, A. and Kehoe, A. (eds.). (2006). The Changing Face of Corpus Linguistics. Amsterdam & New York: Rodopi.

Schmied, J., and Haase, C. (2002). Grammatik lernen im Internet: Die Chemnitz Internet Grammar. In: Keitel, E., Boehnke, K., and Wenz, K. (eds.), Neue Medien im Alltag: Nutzung, Vernetzung, Interaktion. Lengerich, Berlin: Pabst Science Publishers.

Sinclair, J. (2004). How to use corpora in language teaching. Amsterdam: John Benjamins.

Swales, J.M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Yoon, H., and Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. Journal of Second-Language Writing 13, 257-283.