Conceptualizing Causation in Novice and Intermediate Academic Writing

Dr. Christoph Haase
English Department
Purkinye University (UJEP)
Ceske Mladeze 8
40001 Usti nad Labem
Czech Republic

Abstract

This contribution intends to show the confluence of two strands of research that bring together the teaching of linguistics with academic writing and which show that the linguistic perspective can be illuminated by the academic practice. The corpus compiled to this end will be described as well as a number of linguistic features in student writing, usually from students who have never written an academic paper in their lives. The feature of causation will be isolated and described with data from standard corpora. In the empirical part, the conceptualization of causation in student writing will be examined. We show marked differences in the conceptualization based on parameters such as language proficiency.

Keywords: Academic Writing, English for Academic Purposes, Corpora, Causation, Student writing

1. Introduction

In the teaching of academic writing for English language majors, initially, when the students lay eyes on texts of research journal caliber they occasionally experience anxiety and need to be informed that linguistics is in fact not such a hard science and also that all academic matters can be a somewhat simplified in the text type of popular academic writing – a text type that is overall graspable even for non-experts (cf. Hyland & Shaw, 2016).

0012AX Any attempt to compute the uncomputable or to decide the undecidable is without doubt challenging, but hardly new (see, for example, Marxen and Buntrock [25], Stewart [33], Casti [11]). This paper describes a hybrid procedure (which combines Java programming and mathematical proofs) for computing the exact values of the first 64 bits of a concrete Chaitin Omega number, U , the halting probability of the universal Chaitin (self-delimiting Turing) machine U, see [13]. Note that any Omega number is not only uncomputable, but random, making the computing task even more demanding. Computing lower bounds for U is not difficult: we just generate more and more halting programs.

The point of the first examples is that in fact the same content can be linguistically transported in different ways and that a large part of the argumentation lies in the language used.

0012NS In a paper to be published in the inaugural edition of MIT's new journal Quantum Information Processing, Calude and Pavlov have shown that a superposition of an infinite number of energy states would allow a quantum computer to do things no classical computer can ever manage—almost like running "forever" in a finite time.This leap means that a quantum computer can overcome Turing's most famous barrier to computing power: the "halting problem". Given any computer program and an input, can a Turing machine tell in advance whether that program will eventually halt or grind away forever? Turing himself proved the answer is no—the program might stop after a couple of days, or a billion years, but the only way to find out is to run it and wait. That might seem no more than a curiosity, but getting round this barrier would be a gigantic breakthrough for science, solving many important questions in maths and physics.

This leads to ideas what makes one text difficult and what makes the other text not. The above examples are taken from a parallel science English corpus called SPACE (for: Scientific and Popular Academic Corpus of English, see Haase 2014, 2017). From this comparison, students can get a vague initial idea on what academic writing in English actually tries to accomplish. The discussed SPACE corpus at present has more than 1.5 million words mainly from the physical sciences and from the biosciences and it has different sources mainly in the preprint server *arXiv* and the Proceedings of the National Academy of Sciences *PNAS*, the latter due to its public and free availability. *ArXiv* is publicly available as well. The corpus has a very special second component because a parallel and corresponding variant to the science papers has been collected from popular science journals. Originals research is written down by science journalists in publications like the *New Scientist*. The process is that original science papers are written by scientists and read and summarized by popular science journalists who conceptualize predominantly in visual and functional analogs.

For example in popularized thinking, electricity is something that we cannot be perceived directly but we can use water pressure as an analog. Anybody who is not a physicist but with general purpose academic education can understand the analog while the original text in its abstractness and highly specialized conceptualizations cannot be understood by non-experts.

This contribution focuses on the uses of linguistic causation and its conceptualization in novice and intermediate science texts. Science writing starts out with the observation that in science causation is needed as a paradigm of linguistic expressions for cause-effect relationships. Cause-effect relationships are the core of the hard sciences and occur when experimental results are analyzed and variables are aligned. Further, the math behind it is established in such a way that what is found to be aligned is a natural law, a stable and repeatable correlation of events A and Bso that if A happens, B will happen and if A would not have happened, B would not have happened in close spatiotemporal contiguity. Often, theoretical researchers also find truth in a certain elegance of their formulae. In sum, the interface between scientific discovery and language persists in the following:

A cause-effect relationship in research holds on basis of

- experimental results
- proof of a mathematical theorem
- “plausibility” / “elegance”

In the philosophy of science, causation is sometimes referred to as the “Cement of the universe” (see Mackie`s eponymous title, Mackie 1980) or the fuel of our mind (Hofstadter & Sander 2013).

- cause-effect relationship in language is established on the basis of
- morpho-syntactic demands of the language used
- a variety mainly verbal collocations to express it

In the following paragraph, the linguistic expressions of causation and their conceptualization are examined.

2. The X Cause Y construction

Overview

Among the linguistic means to establish the cause-effect relationship are the causative verbs *make, let, help* or *get*, all establishing either a direct link (*make the surface hot, get the machine started*), connecting causer and caused, or an indirect link (*make John leave, help us finish the story*) connecting causer, causee, and caused. Further, the class of resultative verbs add an end result (*wipe the surface clean*), thus extending the complementation frame of the verb (Haase 2007). A simple and direct way is the X CAUSE Y construction that actually involves the verb *to cause*.

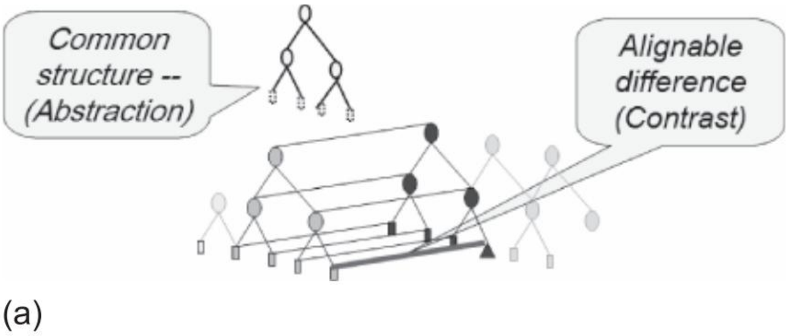
This can be exemplified in its passive (more frequent) as well as active voice in academic texts:
0038AX Although the interaction of a fully ionized and a weakly ionized gas is very complex, an important characteristic can be identified - the generation of magnetic fields caused by relative plasma- neutral gas shear flows. It has been shown (Huba & Fedder 1993) that this process operates

And

0044AX It may lead to step-like transition processes when a small variation of input parameters causes a catastrophic-like transition of the whole system. For example, cloudiness formation caused by the vapor condensation in presence of ions and radicals generated by cosmic rays is very sensitive to vapor concentration

With X being the cause and Y being the effect, the temporal sequence is only retained in active voice while it is reversed in the passive (Kulikov 2006). A less direct conceptualization is therefore the temporally incongruent placement of event B before A in the clause (B *was/is caused by* A). The linguistic analogs for the causes and effects represent mappings that allow insight into the conceptualization of the analogs on the side of the authors. Scientific discovery and analytical thinking are linked very closely the structure mapping approach by Gentner (Gentner 2008) in which the source domain creates relationships between the causal elements and out of these usually easily perceived relationships, an analogue relationship is computed on the target side (Fig. 1).

Fig. 1. Mapping function, cf. Christie &Gentner 2010: 265


Common structure -- (Abstraction)
Alignable difference (Contrast)
(a)

The two levels of this are the computational level which is feasible because they are governed by higher order relationships (usually abstract and mathematical). Visual analogs like comparing DNA to long strings are replaced in the cause-effect relationship by functional analogs which are systematic. Looking at analogy in Academic writing we find that the mapping is constrained by the knowledge of the observer/author but also by the knowledge of the reader and in a higher-order step by the knowledge of the author about the knowledge of the reader (the most complicated to model). This in consequence leads to the emergence of genre types, a consequence of shared and mutual knowledge as shown by the difference between the academic-science and the popular-science texts. Thus, establishing analogical mappings is informative not only about the knowledge of the author but at the same time of the genre conventions (Swales 1993).

2.2 Baseline data on the cause construction

The verb *to cause* was selected as the most unambiguous way to map causes onto effects in language. It also provides a well-entrenched lexical field which enables simple quantification in tagged corpora. When looking at real science texts we can go beyond the impressionistic reading of the texts because it can all be backed up by corpus data together with the queried verb to cause. Looking at their collocates we can show how causes are mapped onto effects and demonstrate this in different text types of the academic genre. In previous studies, especially Haase 2017, it had been expected that *to cause* is dominant in the physical sciences when compared with the biological sciences (a dominance of the parameter of research discipline) and that *to cause* is dominant in academic science writing in comparison with popular science writing (a dominance of the genre/register parameter). This is motivated by the fact that for the physical sciences mathematical formulæ are examples of strict causation as the underlying principle while causation emerges in the biological sciences out of empirical observation. (We skip elucidations her on the social sciences where correlation seems the maximum that can be achieved). For the genre parameter this means that academic science writing is based on logicwhile popular science writing is based on temporal sequence (Haase 2010).

The distribution of *to cause has been* discussed in Haase 2017 according to the variables of research discipline, text type, syntactic environment and domain attribution (concerning the mapping function between source and target domain), carried out with a number of standard corpora as well a the custom-made SPACE corpus. The natural-science core of the corpus was used for the study under exclusion of medicine and psychology so that the search basis was 625,288 tokens. In the inter-corporeal comparison, the differences to the other standard corpora is drastic, see table 1.

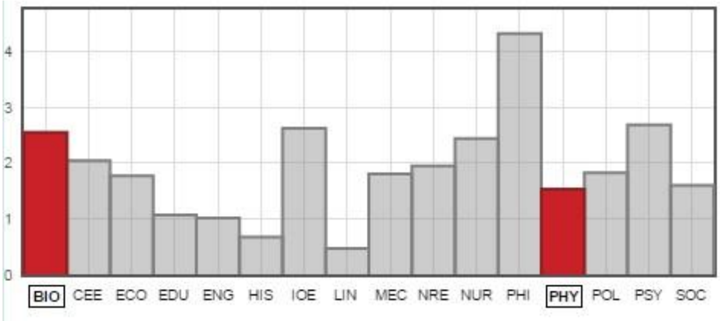
Table 1. Benchmarks of the distribution of to cause

Corpus	f_V	f_V **per** 10^6	**Ac** f_V	**Ac** f_V**per** 10^6
BNC	5,667	58.87	1,260	82.18
COCA	24,282	52.29	5,574	61.21
MICUSP (phys& bio)	1,061	204.04		
SPACE	215	343.84		

The first two corpora given are standard corpora, the British National Corpus and the Corpus of Contemporary American English (COCA). Both are large (100 mio words for the BNC and 560 mio words for COCA) and only have segments devoted to academic texts. MICUSP is the Michigan Corpus of Upper-level Student Papers with a size of 2.5 mio words. Finally, the aforementioned SPACE corpus is 1.5 mio words in size but for this study, only a segment of 0.6 mio words was used. As can be seen immediately, the distribution of to cause in general English (academic as well as non-academic texts) is more or less the same for BNC and COCA (BNC: ca. 59 times per 1 mio words, COCA 52) while the academia-only corpora MICUSP and

SPACE show highly divergent figures for the verb with 204 and 343 occurrences per 1 mio words). Therefore, overall and in academic subcorpora has the British National Corpus the overall lowest figures and in the academic segments the second lowest. This may be due to the fact that what is called academic here are collected texts from political science and law. In their figures, COCA and MICUSP with intermediate to upper intermediate student papers (Römer & Swales 2010) resemble popular science writing. Further it can be attested that the BNC has the overall relatively low frequency of 5,667 verbal occurrences (out of 12,889 token hits for *cause* as a lexical item which includes nominal and informal *'cause* from the spoken component, the latter two have been eliminated from the data). In COCA, the frequency for the academic sections is even lower and only for MICUSP and SPACE the dimensions are in the same neighborhood although the figures for SPACE are 1.5 times higher. As this study, initially published in Haase 2017 considers all inflected forms of to cause, the tagged version of the corpora were used. Here it shows that the difference in tagging methods employed may be responsible for a possible error. BNC has been professionally tagged with the CLAWS tagged which provides considerable accuracy. SPACE has been tagged with TreeTagger, a robust and free method which may contribute to a higher error margin and due to the nature of the project the accuracy could not in all cases be completely verified by human users.

Fig. 2. Distribution of *to cause* in MICUSP proportional in biosciences and physics



To achieve comparability of the two academia-only corpora MICUSP and SPACE, the figures in MICUSP were broken down into the compiled sections and generated with the web interface (http://micase.elicorpora.info/). Here, the MICUSP frequency represents the mean average of two different disciplines, biosciences (BIO) and physics (PHY) shown in Fig. 2. As can be seen, in MICUSP, the biosciences section has a higher proportion than the physics section. Remarkable are the two outliers philosophy and linguistics (PHI & LIN) with philosophy leading the field and linguistics with the lowest counts of *to cause* of all academic disciplines.

The substantial gap between the academic BNC and the other academic corpora can be explained by the fact that the academic texts in the BNC section are not natural-science texts but mainly law texts and in MICASE the count is also lower because novice authors like students do not rely so much on the verb *to cause.* MICASE uses a considerable spread of the verb *to cause* in different domains, as evidenced by the heterogeneous bars in the chart in Fig. 2. For example in the biological sciences the count is 250, in physics 200 and 404 in philosophy.

We can complement the picture by adding the popular-science sections thus fully employing the 1.5 mio words of the SPACE corpus and generate the following numbers (Tab. 2).

Tab. 2. Academic and popular-academic distribution

genre/discipline	**Biosciences**	**Physics**	**total**
Popular	578.31	652.42	615.35
Academic	471.94	159.25	315.59
***f*$_V$ per 10^6**	525.12	392.33	343.84

The contingency table shows the disciplines set in relation to the genre (popular versus academic) and shows a discipline dominance for the use which confirms the physical sciences as the strictest conceptualizers of causation not only in their academic but also their popular academic variant. The following paragraph adds the component of second language learner student writing (see also Hardy & Friginal 2016 for a comprehensive overview).

3. Causation in L2 novice and intermediate student writing

3.1 Data set and benchmarks

All data were obtained from the purpose-built corpus of student writing called CUJOE, introduced in Haase 2019.CUJOE stands for Corpus of UJEP Students Of English. It is a learner corpus made of student theses written at Purkinye University (UJEP) over a number of years. All theses are in the linguistics field.

The corpus was modeled after criteria laid out in the ICLE project (Granger et al. 2009) with its base parameters given in Tab. 3.

Tab.3. CUJOE base parameters

Shared features		**Variable features**	
Age	*20-30*	Sex	*75%F, 25%M*
Learning context	*Degree in English*	Mother tongue	*Czech*
Level	*BA, MA*	Region	*Northern Bohemia*
Medium	*Written*	Other foreign languages	*diverse*
Genre	*Linguistics*	Practical experience	*diverse*
Technicality	*Digital*	Topic	*English language linguistics*
		Task setting	*assigned qualification*

Initially compiled to address research criteria for the investigation of the writings of future educators, the corpus has now become a standard resource for L2 academic writing. As a monitor corpus, the project adds texts constantly. Texts are tagged with the TreeTagger tagset and receive meta information in the headers. The basic separation into novice and intermediate authors is made by the compilation in BA (for Bachelor theses) and MA (for Master theses). The sections are shown in Table 4.

Tab. 4. CUJOE section overview

	mean length	**minimum**	**maximum**
all	14,429	3,609	49,665
BA	12,035	3,609	24,822
MA	20,414	8,251	49,665

It is obvious from the table that the MA theses are much larger by about 40%. They also show much more internal heterogeneity with the longest theses six times longer than the shortest. Theses were not chosen for quality (although all passed their defenses) but instead for representativity. In this, they show obvious hallmarks of student writing if their lexical diversity is investigated. For this, a simple measure is the type-token ratio TTR, see Table 5.

Tab.5.Word counts and TTR in CUJOE

	tokens	**types**	**TTR**
all	488,417	24,665	0.0505
BA	297,958	18,978	0.063694
MA	190,459	13,070	0.068624

Interestingly, the type-token ratio for intermediate writers is not considerably higher than for novice writers, probably indicating that the lexical spread is not much different, given that all writers are non-native and the discipline (linguistics) for all collected texts is the same (see also Biber & Gray 2010).

3.2 Causation in CUJOE

The procedure follows the previously carried out study on the different academic corpora in §2. The verb *to cause* was selected for providing the most direct mapping of causes and effects and thus allowing insight into the conceptualization of the authors. The following examples are extracted for BA and MA, the first representing active voice use, the second, passive.

CUJOE023 Bilingual parents are sometimes worried about the possibility of their child to not be aware of two different patterns of language because it might cause some problems with language acquisition of the child in general.

CUJOE021 the correct answerers prevailed. The exception was question number two, but this was probably caused by the terminology of this word, and it is not much used among the tested people.

The passive to active ratio for BA was 0.32 which is surprising as the use of *to cause* is by far predominantly an active one.

The examples for the intermediate writers (MA) are:

CUJOE103 allow (Y as in yes) for lexical change. This can cause recognition of idioms to be slower. Table 6. Grammatical change. Table 6 shows what kind of grammatical change if any the idioms allow for.

CUJOE107 in both SAT and ESAT, but they are present in both registers. Their low usage rate is caused by high articulation and speaking rates of ESAT commentary. Heavy modifiers are too difficult to implement into ESAT seamlessly and the

Equally surprising here is the passive to active ratio of 0.47, more passives than the novice writers but again, active voice dominates the conceptualization as more direct, temporally congruent. The total data and its proportions are given in Table 6.

Tab. 6. CUJOE data for *to cause* in both sections

	BA	**MA**
tokens	297,958	190,459
caus*	108	47
per 1 mio	362.47	246.77

The collocate words in both sections shed some light on the causes and effects but the size of the corpus prevents them from being truly illuminating. The revealing collocates appear with a frequency of one so they are missing from the list in Table 7. There is no overlap between the sections but due to the smaller MA section, the more specialized texts here show also (like their novice counterpart) only extremely general items like *error* or *word*.

Tab. 7. Collocates of *to cause*

	CUJOE – BA		**CUJOE MA**
freq.	collocate	freq.	collocate
9	*language, error/s*	5	*differences*
5	*changes*	4	*confusion, word/s*
4	*speakers*	3	*problems, preposition*
3	*question, OE, misunderstanding*	2	*result, problem, difference, CzEnglish, anxiety*
2	*violation, threat, separation, pseudo-tuberculosis, problems, number, nature, NAE, maxim, infection, human, gender, frequency, English, effect, communication, bacilli, aspects, ash, ambiguity*		

In the end, the low specification of the lexico-semantic nature of these items underline that the investigated texts are representative of second language learner output in the academic field. Their differentiation into novice and intermediate shows in both sections beginner-level conceptualization.

4. Conclusion

The identified distribution of the verb *to cause* in the two different sections of CUJOE allows an attribution of causation under the parameter of proficiency only insofar as the data basis is still relatively small and the overall number of hits - though comparable - blurs the hypothesized true relationships. the assumption that novice writers, second language or otherwise, show conceptualization behavior akin to that of popular-science writers while higher proficiency writers would be more resemblent of academic-science authors could not be substantiated although trends in the data point in that direction. The carried out study shows predominantly that such behavior can be attributed to conceptualization differences (cf. the work on the standard corpora) and that simple lexicostatistic methods can shed light on such behavior. It is envisaged to enlarge the data base and extend the investigation to all other forms of causation, not only the lexical field of *caus** and its cognates but also to causative verbs and morphological causatives.

5. References

Biber, D. & B. Gray (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. Journal of English for Academic Purposes 9, 2-20.

Christie, S. & Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by
Structural Alignment. Journal of Cognition and Development, 11(3), 356-373.

Gentner, D. & Bowdle, B. (2008). Metaphor as structure-mapping. In: The Cambridge handbook of metaphor and thought, Edited by Gibbs, R.W. Cambridge: CUP, 109-128.

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. (Eds.). (2009). International Corpus of Learner English. Version 2. Louvain: UCL Presses Universitaires de Louvain.

Haase, C. (2019). CUJOE - A New Academic Learner Corpus of English. In: Haase, C. & N. Orlova (Eds.). (2019) English Language Teaching through the Lens of Experience. Newcastle: Cambridge Scholars, 129-134.

Haase, C. (2017). Analogical mapping of domains in cause-effect representations. A comparison of different science text types. Studia Anglica VII. Annales Universitatis Paedagogicae Cracoviensis 226, 106-120.

Haase, C. (2014). A register approach to analogy in science texts: Popular vs. specialized text types. Discourse and Interaction, 7(1) (2014) 33-48.

Haase, C. (2010). Verb classes and the grammaticalization of causativity in discourse. In: Witczak-Plisiecka, I. (ed.) Pragmatic Perspectives on Language and Linguistics 2009. Vol.1: Speech Actions in Theory and Applied Studies. Sheffield: Cambridge Scholars.

Haase, C. (2009). Resultative vs. causative event framing: Description, modeling, problems. In: Povolna, R., & Dontcheva-Navratilova, O.(eds.), Discourse and interaction 1 (2009) 33-47.

Haase, C. (2007). A Crosslinguistic View on Causativity: Causer Neglect. In: Povolna, R., & Dontcheva-Navratilova, O. (eds.), Discourse and Interaction 2. Brno Seminar on Linguistic Studies in English: Proceedings. Brno: Masaryk University, 57-70.

Hardy, J. A. & E. Friginal (2016). Genre variation in student writing: A multi-dimensional analysis. Journal of English for Academic Purposes 22:119-131.

Hofstadter, D.R. & Sander, E. (2013). The forgotten fuel of our minds. New Scientist 218, 2915, 30-33.

Hyland, K. (2016). General and Specific EAP. In Hyland, K. & Shaw, P. (Eds.), The Routledge Handbook of English for Academic Purposes. London and New York: Routledge, 17-29.

Kulikov, L. (2006). Passive in Indo-European. Reconstructing the early Vedic passive paradigm. In: Passivization and Typology. Form and function, Edited by Abraham, W. &Leisiö, L. Amsterdam/Philadelphia : Benjamins, 62-81.

Mackie, J.L. (1980). The Cement of the Universe: A Study of Causation. Clarendon Library of Logic and Philosophy. Oxford: OUP.

Römer, U. & Swales, J. (2010). The Michigan Corpus of Upper-level Student Papers (MICUSP). Journal of English for Academic Purposes 9, 249.

Swales, J. (1993). Genre and Engagement. Revue belge de philologie et d'histoire, 71, 3, 687-698.

Online Sources

International Corpus of Learner English (ICLE) project webpage:
http://fltr.ucl.ac.be/fltr/germ/etan/cecl/
Michigan Corpus of Upper-level Student Papers
http://micase.elicorpora.info/