CAUSAL CHAINING AND THE ACTIVE/PASSIVE RATION IN SCIENCE TEXTS

Christoph Haase

Abstract
In this contribution, a number of constructions of the type X CAUSE Y (and its morpho-syntactic variants) are investigated under the perspective that the causer and the caused (and perhaps the causee) stand in a causal relationship but also in a relationship of different degrees of agency. The most obvious syntactic signature of this is the difference between active and passive sentences. It also offers the advantage to be quantified and examined statistically. For this end, a corpus of academic texts from the natural sciences was queried and the frequencies and domains analysed. The findings may help to improve understanding the rationale behind causal chaining and facilitate cause-effect comprehension for the reader due to inferencing in the ‘correct’ or reversed order. They further demonstrate that the scientific process (which involves the finding of natural laws by means of an establishment of cause-effect relationships) is gradually different in ‘softer’ sciences (which often operate on correlation instead of causation).

Key words
causation, agency, active-passive, corpus linguistics, English for Academic Purposes, EAP

1 Introduction
While the cause-effect relationship in the mind of the researcher holds on the basis of given evidence – be it from experimental results, proof of a mathematical theorem or by a more vaguely defined ‘plausibility’ or ‘elegance’, the linguistic reification is necessarily tied to the morpho-syntactic demands of the language used. The causal chaining (i.e. the establishing of a causal relationship due to observation or theory) therefore relies on the demands of the lingua franca of academia, English. The linguistics of causation (in English but also in other natural languages and in the language of formal logic), however, is widely studied (cf. Haase 2006, 2010) and the realizations accepted in the philosophy of science (beginning with Mackie 1980). However, a long-standing question has been which role the protagonists play in the process of knowledge-generation in the sciences and its subsequent formulation in science writing. The protagonists of the cause-effect relationship therefore are the causer and the caused and in some cases the causee (an intermediate protagonist who or which is coerced into the causal chain by the causer). What is called protagonists here of course can be
the researchers themselves (who have assumingly a free will and ‘intentionality’) but also the objects of research which are ‘active’ but without intentionality, cf. “Last year’s heat wave in Europe caused at least 20,000 deaths” (SPACE corpus 0097NS). This has been expressed by Flaherty as a central issue: “By raising the specter of intentionality, we can glimpse the alternation between yin and yang in the human experience of causality: sometimes things happen to us; sometimes we make things happen” (Flaherty 2011: 3).

In a nutshell, the particular features of generating knowledge via a number of syntactic constructions can be studied by looking at the most unambiguous of all, the X CAUSE Y construction and its two syntactic realizations as active and passive. Causatives are grammaticalised from intransitives which do receive passivisation except if they appear as anticausatives (“The bottle broke”) and the agent is irrelevant and where an autocausative interpretation (“The bottle broke itself”) is nonsensical. Therefore the construction does not allow “The bottle was broken by itself”) (for a similar example, cf. Wiemer 2006: 290). The agency of the causer appears in unmitigated form in all active constructions. The causer is syntactically optional in all passive constructions and it will be illuminating to investigate whether science writing on different levels supplies causers in all cases. Authors, however, are constrained by academic style-guides which teach passive styles, mainly to eliminate the subjective element (Hyland 2006: 14) but they are more frequent in introduction and discussion (ibid.: 141). However, the CAUSE construction is a special case as the causers and thus agents are only in the rarest cases free-will subjects (aka: humans). Still, the “passive pressure” prevails (Canagarajah 2002: 134).

The corpus used in this study contains a large number of different causative constructions in texts from diverse natural sciences and in different registers (academic and popular-science). This is illustrated in the following example from an academic and a popular-science publication:

(1) 0052PN …with lower antioxidant status and heightened mitochondrial oxidant production during aging. Addition of ALCARLA to the diet of old rats caused a significant decline in appearance of oxidants to 2,801.79 fluorescence units/min per mol O2 consumed per 106 cells, which was ...

(2) 0059NS … of the oceans is leading to disaster. Salmon fisheries off the north-west coast have collapsed, there is a dead zone in the Gulf of Mexico caused by polluted run-off, and foreign species arriving in the ballast water of cargo ships are disrupting local ecosystems. All these problems ...
In (1) the causer (“addition of ALCARLA”) has agency, is subject, has primacy (due to its initial sentence position), semantic prominence etc. In (2), the causer (“polluted run-off”) in comparison is much less prominent as an object of the by-phrase.

A main purpose of this paper is to investigate whether the difference in style between academic registers is relevant in the attribution of agency to the causers in an active-voice sentence, or in other words, if a less formal style, as found in popular science publications, leads to a conceptualization of scientific facts as – in Flaherty’s words – us making them happen or if they are happening to us (cf. above). We propose with Swales that in academic writing, the attribution of agency is further constrained not only by the knowledge of the observer (and in our case the author of the study or by the experimenter) but also by the knowledge of the reader and, thirdly and most complicatedly, by the knowledge of the author about the knowledge of the reader (or his/her theory of that knowledge). The emergence of genre types (or: academic registers such as specialized-academic and popular-academic) is therefore a consequence of this shared and mutual knowledge and the attribution of agency is informative not only about the knowledge of the author but at the same time of the genre conventions. At the same time, the genres remain abstractions (Swales 1993: 688). While causal chaining in language represent the backbone of human understanding of nature in the natural sciences, causation in language actually represents a reflection of nature in the sense that the identified cause-effect relationships need further investigation. The verb to cause has been selected as the most natural way cause and effect can be chained unambiguously. We can see this in the following examples from the corpus (to be described later). The number of examples is limited and in a break-down of the actual figures it can be obtained that cause is much more often used in the natural sciences than in the biological sciences. This may be due to the fact that in the natural sciences causation is often the underlying assumption of mathematical formulae whereas in the biological sciences causal chaining happens due to observation and to the precise recording of results. This means also that for these sciences the application of statistics is virtually the only way to establish cause-effect relationships.

In the next paragraphs, the agency of the causers is studied on the example of the verb CAUSE. It is motivated by two reasons. First, it is the foremost privilege of scientific thought to find causation instead of correlation (a domain of the social sciences) and this is given primacy over pure conjecture. Second, it represents a well-entrenched lexical field which enables a quantitative (and thus statistically sound) look at their occurrences but also allows a qualitative view as well as the number of findings in corpus can then be discussed almost on a case-by-case basis.
2 Preliminaries

2.1 On causation in science writing

The question that is central to this study is whether causation is a real or a represented aspect in academic texts and academic writing. Observing phenomena and connecting these observations and phenomena in order to establish stable and repeatable relationships is the core of science but in the study of EAP (or English for Academic Purposes) the goal is to find ways to systematize the grammaticalisation and lexicalization of these relationships. Therefore, language itself becomes the instrument of reasoning. Studying linguistic features can help improve and sharpen this instrument. Phenomena in nature are principally connected by three different links: The first and weakest link is chance occurrence of two phenomena and it does not hold under repetition. If phenomena occur stably and repeatedly together this is called correlation and it helps to hint at a third phenomenon which causes both. The strongest and most ‘valuable’ link is causation in which the phenomenon is linked to another phenomenon repeatedly and by removing the first, the second disappears. Causation can be supported best with mathematical formulae.

The initial question is whether causation as displayed in academic texts and scientific writing allows for a certain tolerance with the placement of the causers as subjects of active-voice constructions or as objects of the by-phrase in passive-voice constructions. This is evidenced by many texts completely ignoring the user in terms of agency to the attribution of causation, for example here:

(3) 0017AX The rain rate can be calculated instantaneously as \( q(t) = \sum_{i} n_i V_i \). In the time series we investigated, the continuous measurement is averaged over one-minute intervals, leading to one minute temporal resolution. When the signal due to rain becomes indistinguishable from the background noise at the receiver, the rain rate is defined as zero.

The conclusions of the readers are therefore very often not taken directly from the causative verbs or other expressions of causation but because the reader receives the message from involvement markers or commitment markers in these texts. The agents are inferred, supplied by world knowledge or domain knowledge of the reader, in the case of (3), the object of the unexpressed by-phrase is the experimenter.

2.2 Causatives and the verb CAUSE

English causatives as many of the English verbs show alternations in valence but there is no morphological marking of the different valence classes (cf. Blevins
2006: 526). This is especially true with verbs reflecting the pattern of $X \text{ CAUSE } Y$ in a syntactic sense. In an evolutionary view, they are old constructions, making their first appearance in early Vedic (as derived from intransitives, often verbs of perception and consumption (‘see’, ‘know’, ‘drink’), cf. Kulikov 2006: 76). Most of these verbs can show the alternation between causation and inchoation, verbs like sink, break or drown. Many of these verbs can also be used intransitively. The verb cause cannot be used intransitively and it represents the most direct way from cause to effect. It thus deviates from the pattern on mediated causation that involves a cause or causer, the effect, and a causee as a go-between. The pattern is $X \text{ MAKE } Y \text{ CAUSE } Z$ and it was studied in Haase (2009). In the following data discussion, these constructions are investigated in a variety of standard corpora (BNC, COCA, and MICUSP) and subsequently in the SPACE corpus.

3 Materials and data discussion

3.1 The corpus

The queried corpus is called SPACE (for Specialized and Popular Academic Corpus of English) and was compiled on the basis of research papers and their summaries in popular science publications. It has been described in considerable detail in this journal (most recently in Haase 2014); therefore the following facts will be brief.

The SPACE corpus was designed to show marked differences between registers and disciplines. These disciplines persist on the one hand between different fields of research or science and on the other between different text types and an integration of the two different text types exists, one type being the traditional research article written by practitioners and scientists and targeted at other scientists or the peers of the original scientists. The other text type is the type of popular academic writing and we find it in popular science journals like the Scientific American or the New Scientist. The target group here is scientists or academics who are interested in science beyond their own field; thus these articles were not written for the peers of the original researchers but they represent popular summaries of original research. In general, these articles are much shorter than the original research and they may contain references to additional papers or previous research. In the SPACE corpus it was attempted to create a parallel structure in which we have the possibility to compare the specialized and the popular article in parallel. This has been done for the reason of finding linguistic parameters of academic writing under the viewpoint of target audience and text type. The scientific disciplines involved are physics on the one hand and the biosciences on the other. The physics component of the SPACE corpus collects original research and popular summaries in cosmology, particle physics,
quantum physics and astrophysics. All were retrieved from the pre-print server arxiv (arxiv.org). The biological sciences in the corpus involve biochemistry, genetics, molecular biology, microbiology, and partly medicine. This enables us to study several linguistic features, for example, the common syntactic features of academic writing like nominalization and the passive voice but also the use of metaphor and figurative language. Among the pragmatic features to be studied is, for example, the study of commitment markers and other discourse-based moves (as e.g. studied by Flowerdew 2008), which is particularly different because in the original research the researchers stand for the results whereas in the popular summary the author often arrives at an informed conclusion on the basis of the work of others; it is thus second-hand research which changes the viewpoint of the author to the object of research.

3.2 The distribution of CAUSE

In a simple breakdown we discuss the occurrences of CAUSE in the corpus according to the parameters of syntactic environment, text type, research genre and most importantly, domain attribution. While the more technical numbers show frequencies of the different (and differing) uses of to cause, they make sense only in comparison to the competitors of cause, of which there are numerous. They have been discussed, for example, in Haase (2006). For this study, a number of subcomponents have been selected that represent the natural-science core of the corpus, under particular exclusion of medicine and psychology. The basis for all obtained ratios is a word count of 625,288 words. A simple search yields 91 occurrences of cause, which is a meaningless number as it contains all nominal uses but only a fraction of the verbal uses of which caused is of course the most frequent one (137 hits). There are 40 nominal uses – 28 singular and 12 plural.

A look across different corpora shows wildly divergent figures for the occurrences of CAUSE in academic as well as other compiled text genres. The British National Corpus has overall the lowest numbers, also in the academic parts. This may be due to the fact that the academic part mainly consists of text from political science and law. It is still surprising that some of the figures by far are outnumbered by all the other corpora. The other corpora come close to the figures found in the SPACE corpus but do not approach their numbers. This could be explained with the condition that the texts collected in COCA as well as MICUSP do not approach the level of scholarship of the papers collected in the SPACE corpus. COCA collects texts from contemporary American English and only devotes one section to academia while MICUSP in effect is a very small corpus and relies on intermediate to upper intermediate student papers (cf. Römer & Swales 2010).
<table>
<thead>
<tr>
<th>Corpus</th>
<th>$f_v$</th>
<th>$f_v$ per 10^6</th>
<th>Ac $f_v$</th>
<th>Ac $f_v$ per 10^6</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>5,667</td>
<td>58.87</td>
<td>1,260</td>
<td>82.18</td>
</tr>
<tr>
<td>COCA</td>
<td>24,282</td>
<td>52.29</td>
<td>5,574</td>
<td>61.21</td>
</tr>
<tr>
<td>MICUSP (phys &amp; bio)</td>
<td>1,061</td>
<td>204.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPACE</td>
<td>215</td>
<td>343.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Distribution of CAUSE in different corpora

The difference in the corpora can be explained when we look at the breakdown of the data. The BNC has an overall relatively low frequency of 5,667 occurrences of CAUSE as a verb (out of 12,889 hits for cause as a lexical item, including the nominal use and the informal ‘cause from the spoken component). While the frequency in the academic part is considerably higher, we find the frequency (of the verb) $f_v$ per million (10^6) resolved as 82.18. In the COCA corpus the academic frequency of the verb is even lower. Only the figures for MICUSP and the SPACE corpus are roughly in the same dimension. Still, the SPACE corpus has an almost 1.5 times higher frequency than MICUSP. Together with the attempted explanation above, however, we have to consider that the MICUSP frequency is the mean average of two different disciplines, physics and biology. The difference here is that in the biological component of MICUSP the frequency of CAUSE is almost twice the number of the physical frequency. The following graph breaks down the academic disciplines in MICUSP.

![Figure 1: MICUSP proportional distribution of CAUSE in biology and physics (generated with http://micase.elicorpora.info/)](http://micase.elicorpora.info/)
However, the MICUSP corpus interface is relatively intransparent about the actual figures (except for the size of ca. 2.6 mio words, Römer & Swales 2010: 249). The graph therefore shows proportions only. Interesting is the outlier of 410 per million for philosophy and quite surprising the minimal 50 uses of *CAUSE* per million words for linguistics.

As for the large deviations in general between the corpora, another issue could be the part-of-speech tagging. A considerable margin of errors may actually be introduced via tagging errors. While the BNC has been tagged with the CLAWS tagset and seems considerably accurate, SPACE was tagged with Tree Tagger which accrues a higher error margin and has not completely been verified by human users. The following table displays the distribution of all uses of the verbal paradigm of *CAUSE*, as identified in the tagged version of the corpus.

<table>
<thead>
<tr>
<th>Tag</th>
<th>AcBio</th>
<th>AcPhys</th>
<th>PopBio</th>
<th>PopPhys</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>VVN</td>
<td>74</td>
<td>24</td>
<td>3</td>
<td>6</td>
<td>107</td>
</tr>
<tr>
<td>VV</td>
<td>12</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>34</td>
</tr>
<tr>
<td>VVZ</td>
<td>14</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>VVG</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>VVD</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>VVP</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>total (f_v)</td>
<td>126</td>
<td>46</td>
<td>18</td>
<td>25</td>
<td>215</td>
</tr>
<tr>
<td>Tokens</td>
<td>266,982</td>
<td>288,862</td>
<td>31,125</td>
<td>38,319</td>
<td>625,288</td>
</tr>
<tr>
<td>(f_v) per 10⁶</td>
<td>471.94</td>
<td>159.25</td>
<td>578.31</td>
<td>652.42</td>
<td>343.84</td>
</tr>
</tbody>
</table>

Table 2: Overview of the different verbal uses of *CAUSE* in the four SPACE components

The distribution in itself offers few surprises: VVN emerge at the top as a passive-dominated style favours participles, VV (infinitives and non-finite forms) account for less than one third, VVZ (simple present forms/singular) are more frequent than their past-tense counterparts (VVP) and participles (VVD), VVG (-ing forms) assume a medial position. More interesting is the share per million words, especially compared with the other corpora. While the same ranking in specialized biology and physics applies as in MICUSP, the difference within SPACE is quite striking: 472 academic biosciences vs. 159 academic physics, cf. the respective text samples below:
Increasing plant diversity beginning in the Silurian (425 million years ago) led to increasing weathering of rocks that had two effects: atmospheric CO2 levels decreased, causing a decrease in carbon isotope fractionation in marine deposits;

... the number of systems that change from \(sA = +1\) to \(sA = -1\) is unequal to the number that change from \(sA = -1\) to \(sA = +1\), causing an imbalance that changes the outcome ratios at \(A\). In other words, in general the statistical distribution of outcomes at \(A\) is altered by the distant shift ...

Overall, the use in the popular text-types is much higher with 578 in popular biosciences to a colossal 652 in popular physics:

The unpredictable accelerations caused by these spurts are at least 10 times as big as the Pioneer acceleration, and make it impossible to measure the effect.

To insert the gene, the team first incorporated it into an adenovirus similar to the one that causes colds. They also plucked out genes that enable the virus to replicate, so that it would load its genetic cargo into mouse cells without infecting the rest of the animal.

Thus in conclusion, the text types compiled in the SPACE corpus overall rely on CAUSE and use it in larger quantity than comparable corpora although no means of comparison concerning morpho-syntactic variability of use can be established. The focus therefore needs to be narrowed to the agency as expressed in active and passive forms.

3.3 Distribution of CAUSE in active and passive

Out of the verbal paradigm of CAUSE, only the –ed/–en forms were selected identified by their past-tense tags (VVP) and participle tags (VVD). The domains were identified in both directions for the active sentences (\(X\) CAUSED \(Y\)) and for the passive ones (\(Y\) was CAUSED by \(X\)). In the following, the findings will be discussed in the individual registers together with one example each for the direction (cause before effect in active, effect before cause in passive).

Popular biosciences

...injury was probably CAUSED by a blow ...

...externally applied field CAUSED cultured chick neurons to grow ...

...
Table 3: Directional distribution for CAUSED in popular bioscience texts

In popular bioscience, the ratios of both directions are nearly the same size which is surprising as this is the only register in which this is the case, in fact there is a slight majority of the X CAUSED Y pattern. The active/passive ratio is 1.22. The comparatively high indication of actives can be explained first by the popular nature of the texts (an argument which could be made also for the popular physics texts but it turns out it is less important there) but also by the concreteness of the domains of cause and effect. Concrete causers fit an active syntax better than the highly abstract causers in physics where no agency can be attributed.

Table 4: Directional distribution for CAUSED in academic bioscience texts

By far most CAUSED appearances come in passive syntax with only nine out of 82 occurrences showing the effect preceding the cause. The active/passive ratio is 0.12.

Popular physics
(12) 0003NS *The unpredictable accelerations CAUSED by these spurts ...*
(13) 0004NS *Interactions then rapidly CAUSED it to relax into the special “equilibrium state” ...*
CAUSAL CHAINING AND THE ACTIVE/PASSIVE RATION IN SCIENCE TEXTS

<table>
<thead>
<tr>
<th>Direction</th>
<th>$f_{absolute}$ (relative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y (effect) was CAUSED by X (cause)</td>
<td>6 (75%)</td>
</tr>
<tr>
<td>X (cause) CAUSED Y (effect)</td>
<td>2 (25%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8</strong></td>
</tr>
</tbody>
</table>

Table 5: Directional distribution for CAUSED in popular physics texts

The trend indicated on the example of the popular biosciences can be observed here as well, only less defined. The active/passive ratio is 0.33. Still, 25% active CAUSE sentences is an unusually high amount.

Academic physics

(14) 0003AX... the earth-rotation vector is CAUSED by the gravitational torque ...

(15) 0003AX Each manoeuvre CAUSED a change in spacecraft spin ...

<table>
<thead>
<tr>
<th>Direction</th>
<th>$f_{absolute}$ (relative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y (effect) was CAUSED by X (cause)</td>
<td>33 (87%)</td>
</tr>
<tr>
<td>X (cause) CAUSED Y (effect)</td>
<td>5 (13%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>38</strong></td>
</tr>
</tbody>
</table>

Table 6: Directional distribution for CAUSED in academic physics texts

For both physics registers, passive syntax places the effects before the causes in the order of mention. The active/passive ratio is 0.15. While this is to be expected in any register of science writing the much more balanced (technically speaking even reversed) order for popular bioscience could be explained with the nature of the field that in itself is not very abstract (very concrete processes at cellular or molecular level are described) but their explanation lends itself more to a step-by-step description.
The ratios are summarized in the table below:

<table>
<thead>
<tr>
<th>subcorpus codes</th>
<th>pop-bio</th>
<th>ac-bio</th>
<th>pop-phys</th>
<th>ac-phys</th>
</tr>
</thead>
<tbody>
<tr>
<td>ratio</td>
<td>1.22</td>
<td>0.12</td>
<td>0.33</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Table 7: Active/passive ratios for all SPACE dimensions**

The high ratio for popular biosciences and the much lower ratio for academic biosciences (by a factor 10!) represent the outliers among the corpus dimensions. They show that this parameter has considerable influence in the attribution of agency to causes in subject position in popular registers although the value for popular physics is more similar to both academic registers.

**4 Conclusion**

The causal link at the centre of scientific discovery was examined in the example of the verb *CAUSE* which could be obtained in rich diversity in the corpus. This richness was juxtaposed with its more restricted distribution in other comparable corpora. Further, the register parameter could be isolated as important with popular science texts using a larger number of examples from the verbal paradigm of *CAUSE* than the academic texts. We can conclude that all registers of academic writing rely on causal relationships and are only divided by the attribution of agency to causes in subject positions/active sentences. While the use of passives is recommended standard in academic style-guides and practice (the most extreme example section from the corpus yields 89% passives), in the ‘softest’ science investigated (the biosciences) and in its popular variant, actives and passives occur in almost equal numbers. Returning to the initial question whether the scientific process is based on unambiguous causal chaining in the ‘hard’ physical sciences and perhaps to a lesser degree in the ‘softer’ biosciences and if active/passive syntax can play a mitigating role, we can attest that all sciences in all registers use causal chaining as evidenced by the abundance of *CAUSE* constructions across all dimensions. The active/passive distribution, however, represents an interesting yardstick for these dimensions.

**References**


**Christoph Haase** is a Researcher and Lecturer in English Language and Linguistics. With a background in the natural sciences and in English and German linguistics, he is primarily interested in morphosyntactic phenomena of temporality and causation from a cognitive perspective. His other research interests are in first and second language acquisition, English language teaching, and corpus studies in the field of English for academic purposes. After ten years of working at German universities, his academic home is the English Department at the University of J. E. Purkyně, Ústí nad Labem, Czech Republic.

**Address:** Christoph Haase, English Department, Faculty of Education, University J. E. Purkyně (UJEP), České mládeže 8, 400 96 Ústí nad Labem, Czech Republic. [e-mail: christoph.haase@ujep.cz]