Causation in a Learner Corpus of Hungarian-English Summaries of Disease Descriptions

Anita HEGEDŰS
Department of Languages for Biomedical Purposes, Medical School,
University of Pécs

Abstract*: A learner corpus is a collection of language data stored electronically. This paper undertakes to investigate how Hungarian L2 learners of English use causal relationships in a learner corpus of summaries of disease descriptions submitted by Hungarian medical students at level B2 (upper intermediate). The 'UPMS Learner Corpus of Summaries' has been expanding since March, 2022. The data were processed using Sketch Engine. At present, the corpus contains 108 summaries, 24,321 tokens, 21,397 words and 1,244 sentences. In the corpus, the different forms of the lexical item 'cause' were found to be by far the most common ways of expressing causal relations. It was the 29th most common lexical item and the 10th most common content word in the corpus. Other common ways of expressing this function were the lexical items 'result' and 'reason', with 17 and 14 appearances, respectively. However, all the occurrences of 'reason' were used incorrectly in semantic terms and in 8 cases there was a syntactic mistake, too. The implication of the study for teaching medical English is that more emphasis should be put on the semantic distinction between lexical items expressing causation as well as on their syntactic patterns.*

Keywords*: causation; learner corpora; summary; disease descriptions; Sketch Engine;*

Introduction

Several definitions of causation, also referred to as causality or causal relations, have been proposed in linguistics and philosophy. Drury et al. (2022) define causation as a temporal relationship where a cause event activates an effect event at a subsequent time. Vendler (1967) emphasizes the dependency of the cause and effect on each other, i.e. an effect must always be preceded by a cause, it cannot arise without a cause. Hume (quoted in Holland 1986) puts forward three conditions for causality: (1) the cause must precede the effect in time, (2) there is spatial and temporal contact between the cause and the effect, and (3) the cause and the effect either co-occur or neither of them occurs. Mill (quoted in Holland 1986) also postulates three conditions for causality: (1) the cause must precede the effect, (2) there is association between the cause and the effect, and (3) lack of other possible causes. Altenberg (1984) describes three features of causative relations: (1) it involves a binary relationship, (2) it reveals the type of relationship between

the two members of the relation, and (3) it pinpoints the two members of the relationship in a logical sequence. There is agreement between the different approaches that a causal relationship exists between certain events, facts and objects (Vendler 1967) and that the cause event must precede the effect event (de Spinoza 1996; Koo et al. 2002).

Blanco et al. (2008) distinguishes between three subtypes of causation: consequence, reason and condition. Consequence applies if the effect results indirectly from the cause, reason refers to causations arising from decisions, feelings or beliefs, while condition applies if the cause is hypothetical. Akkasi and Moens (2021) highlight three distinctions of causal relationships: explicit or implicit, ambiguous or unambiguous and marked or unmarked causal relations. In explicit relations, both the cause and the effect are expressed in an explicit way. In contrast, in implicit relations the causal relationship is only implied. In ambiguous causal relations, the connection between the cause and effect is expressed with ambiguous words in terms of causality (e.g. *after, before*), whereas unambiguous relationships are denoted with unambiguous words indicating causality (e.g. *cause*). A causal relation is marked if the connection between the cause and effect is designated with a definite linguistic unit while in an unmarked causal relation there is no such indicator and the causal relationship can only be inferred from the context. The study intends to investigate explicit, unambiguous and marked causal relationships.

Linguistic units expressing causal relations include causal verbs, which can be assigned to three categories: simple, resultative and instrumental verbs (Girju 2003). Verbs belonging to the first category form the causal link between the cause and the effect event. E.g. 'cause' behaves as a simple causal verb in the relation *'the monthly bleeding can cause iron deficiency*.' Verbs of the second category provide the causal link and also a description of the effect. For instance, the verb 'consume' can be classified as a resultative verb in the sentence *'if large amounts of fatty food are consumed, bloating, a feeling of fullness and possibly diarrhea can occur.*' Instrumental verbs contain both the cause and the causal link. E.g. 'make' acts as an instrumental verb in the causal relation *'certain medicines might make the symptoms worse*.'[1] The study only concerns simple causative verbs. In addition, causal relationships can also be expressed by conjunctions (e.g. *because*) (Lorenz 1999) and prepositional phrases (e.g. *due to*) (Degand 2000). Moreover, nouns and adverbs can also denote causal relationships.

Trimble (1985) underlines the importance of causation in ESP (English for Specific Purposes) communication. He points out that understanding causation is crucial for creating cohesion and coherence in

[1] The examples were retrieved from the UPMS corpus of summaries.

texts. He highlights that causative constructions help link ideas logically in ESP writing, making the text easier to follow and understand.

Akkasi and Moens (2021) highlight the importance of causation in biomedical texts, attributing an essential role to causal relations in diagnosis, pathology and systems biology. The concept of causation was first introduced in medicine by Jakob Henle and was further elaborated by Rober Koch (Evans 1976). In his presentation at the International Congress in Berlin in 1890, he put forward three concepts based on which a causal relationship between a causative microbe and a disease can be established:

> 1. The parasite occurs in every case of the disease in question and under circumstances which can account for the pathological changes and clinical course of the disease.
> 2. It occurs in no other disease as a fortuitous and nonpathogenic parasite.
> 3. After being fully isolated from the body and repeatedly grown in pure culture, it can induce the disease anew. (Evans 177)

While Evans admits that Koch's postulates retain historical significance, however, in his view, they are too simplistic and not in accordance with the development of medicine (Evans 1976). Thagard (1999) argues that a cause of a disease cannot be described based on its correlation with another factor but involve complex causal networks. These networks include alternative causative factors and mechanisms based on which a factor results in a certain effect. In both medicine and linguistics, causal relations can entail a relationship between a cause event and an effect event, which are dependent on each other, and also causal networks. Linguistics provides the semantic and syntactic realizations of these relationships and networks. Moreover, linguistics provides a means for expressing implied and inferred causal relations.

The present study aims to explore how Hungarian medical students, who study English as a second language use causal relations based on a learner corpus of summaries of disease descriptions. Sinclair (2005) defines a corpus as a collection of language texts in an electronic form, which provides a source of data for linguistic research. Several corpora of biomedical texts, which are annotated for causal relations are available in the literature. In the Biocause corpus, texts were manually annotated for causal relations (Mihăilă et al. 2013). Gopalan and Devi (2017) compiled a corpus of abstracts accessed from PubMed to develop a dataset for explicit causal relation extraction. Sharma et al. (2018) complied a corpus from 10,000 PubMed abstracts concerning leukemia to extract causative verbs. Learner corpora are systematic, electronic collections of texts, but the texts are produced by

language learners (Granger 2008). Learner corpora have two main functions: they contribute to second language acquisition by giving a description of interlanguage and they can be employed to develop teaching methods and tools that target the needs of language learners. The present study investigates how learners of medical English express causation and what linguistic units and patterns are employed for this function.

Methods

The 'UPMS Learner Corpus of Summaries' contains summaries of disease descriptions written by Hungarian medical students preparing for the English for Medical Purposes Exam (EMP) at level B2. Two EMP exams are available at Medical School, University of Pécs (UPMS): a state-accredited EMP exam (PROFEX) and an in-house EMP exam. Both exams include a task of writing a summary, which entails summarizing the description of a disease in English based on 10 prompts. The Hungarian source text contains 300-350 words and the summary should contain a minimum of 150 words. The UPMS Learner Corpus of Summaries was compiled of summaries submitted by students as homework in the framework of the preparatory course for the exam. The texts produced by students were entered into the corpus in an unaltered form, i.e. grammatical, semantic or spelling mistakes were not corrected. Examples provided in the study are also the learners' original products without any corrections. The corpus has been expanding since March, 2022. Currently, the corpus contains 108 summaries, 24,321 tokens, 21,397 words and 1,244 sentences. The Sketch Engine program was employed for the analysis. In the Sketch Engine, the wordlist, concordance, thesaurus and word sketch functions were used. In some cases, the qualitative analysis was supplemented by quantitative analysis. The results of the analyses were collated with the Medical Web Corpus to investigate to what extent the learners' use of causative devices compares with that of L1 language users. The Medical Web Corpus contains over 33 million words assembled from texts in English related to the field of medicine.

Results

Similar to most corpus analyses, the most frequent words in the UPMS Learner Corpus of Summaries are function words, i.e. articles, pronouns, conjunctions, the copula and prepositions. The results are depicted in table 1.

word	no. of occurrences
1. the	1,944

2. be	1,093
3. of	859
4. in	588
5. and	456
6. can	423
7. iron	420
8. a	302
9. to	291
10. or	269
11. cause	216
12. deficiency	195
13. it	188
14. by	183
15. case	176
16. with	148
17. disease	148
17. for	148
19. symptom	140
20. not	139

Table 1. The 20 most frequent words in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

The most frequent word expressing causation is 'cause', which is the 11th most common word in the corpus. Following 'iron', 'cause' is the second most frequent content word in the corpus. Of its 216 occurrences, it appears as a verb in 131 instances and as a noun in 85 instances. By way of comparison, 'cause' is also the most frequent causative word in the Medical Web Corpus: it is the 81st most common word with 49,418 occurrences. In the Medical Web

Corpus, it also appears more frequently as a verb than as a noun (29,902 occurrences vs. 19,174 occurrences).

Causative verbs in the corpus

The most frequent verbs expressing causation are displayed in Table 2. The table shows verbs with at least five occurrences.

verb	no. of occurrences
1. cause	131
2. lead	32
3. result	17
4. trigger	6
5. provoke	5

Table 2. The most frequent verbs expressing causation in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

As can be seen in Table 2 above, study participants predominantly used the verb 'cause' to express causality in the corpus. It is the second most common verb in the whole corpus following the copula 'to be', i.e. 'cause' is the most common content verb in the corpus. Out of its 131 occurrences, in 54 instances the verb occurs in the passive. 'Cause' is most commonly premodified by an adverb (in 32 instances), The most typical collocation is 'can also be caused by' (n=10), e.g. *it can also be caused by an incorrect diet*. Other common collocations include 'is usually caused by' (e.g. *iron deficiency, which is usually caused by bleeding*), 'is most often caused by' (e.g. *which is most often caused by bleeding*) and 'can also cause', (e.g. *Air ventilator can also cause barotrauma in the lungs*). 'Lead' is the second most frequent causative verb and the 11th most frequent verb in the corpus. It is followed by the preposition 'to' in 31 cases. It is grammatically incorrect in the only instance where it is not followed by 'to': *that leads lung collapse and immediate dyspnea*. 'Lead' is preceded by the modal auxiliary 'can' in 11 cases, e.g. *organ failures can lead to death*.

The verb 'result' is followed by the preposition 'in' in 9 cases, e.g. *it results in collapse of lungs*. In 5 cases, the verb is followed by the preposition 'from', indicating a reverse direction compared to 'result in': when using 'result from' the effect is placed before the cause in the sentence. By using the phrasal verb 'result from', we focus on the effect in a causal relation, e.g.

atopic dermatitis does not result from an allergic reaction. In three cases the verb occurs without a preposition, which constitutes a grammatically incorrect use, e.g. *which results quick evaporation*. The verbs 'trigger' and 'provoke' are used only in a few cases in the corpus. In these cases, they are mainly used in the active and are modified by a modal auxiliary expressing possibility, e.g. *food allergies may provoke the development*.

'Cause' was also the most frequent causative verb and the 13th most common verb in the Medical Web Corpus with 29,902 occurrences. Similar to the UPMS Learner Corpus of Summaries, 'cause' appeared more commonly in the active than in the passive voice: it occurred in the passive in only 2,761 cases. In contrast, in the Medical Web Corpus 'produce' was the second most frequent causative verb with 20,275 occurrences. Although in some instances 'produce' is used in other meanings, the Thesaurus function of Sketch Engine identified the verb 'cause' as the most typical synonym of 'produce' in the corpus. On the contrary, there was only one occurrence of 'produce' as a synonym of 'cause' in the UPMS Learner Corpus of Summaries: *the keratoconjunctivitis sicca can be produced by disnormal composition of the tears*. 'Lead' co-occurring with the preposition 'to' was the third most common causative verb in the Medical Web Corpus (n= 9,221) followed by 'result' (n= 9,685) in fourth place. 'Result' was followed by the preposition 'in' in 5,009 cases and by the preposition 'from' in 2,658 cases.

Causative nouns in the corpus

Nouns denoting causal relationship in the corpus are depicted in Table 3.

noun	no. of occurrences
1. cause	85
2. reason	22
3. result	12
4. consequence	3

Table 3. Nouns expressing causation in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

'Cause' was the most frequent noun expressing causality and the 12th most frequent noun in the corpus. Out of its 85 occurrences, it is preceded by the adjective 'common' in 26 instances and followed by the preposition 'of' in 22 cases, thus, its most typical collocation being 'the most common cause of',

e.g. *the most common cause of iron deficiency is bleeding*. In 4 cases, the noun is followed by the preposition 'for', which is grammatically incorrect, e.g. *there is no cause for pneumothorax*. All 22 occurrences of 'reason' in the corpus are semantically incorrect. According to the Merriam-Webster dictionary, "reason" is a "statement offered in explanation or justification" or a "rational ground or motive" (https://www.merriam-webster.com/dictionary/reason). On the other hand, "cause" "brings about an effect or a result" or provides "the reason for an action or condition" (https://www.merriam-webster.com/dictionary/cause). Therefore, the two nouns cannot be used interchangeably, as 'cause' implies a direct relationship between the cause and the effect, whereas 'reason' provides an explanation for why an event occurred, and the cause and the event may not be directly related. Moreover, 'reason' requires to be followed by the preposition 'for', and, in the corpus, it is followed by the preposition 'of' in 13 instances. Thus, more than half of the uses of the noun 'reason' are both semantically and grammatically incorrect. 'Result' as a noun occurs 12 times in the corpus. In 6 instances, it occurs in the collocation 'as a result of', e.g. *as a result of an allergic reaction*.

In the Medical Web Corpus, 'result' (n= 29,949) was much more common than 'cause' (n= 19,174). In the whole corpus, 'result' was the 19th most common noun. 'Reason' was the third most common causative noun with 8,568 occurrences.

Causal conjunctions, adverbs and prepositional phrases

Conjunctions expressing causation found in the corpus are displayed in Table 4.

conjunction	no. of occurrences
1. because	49
2. since	20
3. as	15

Table 4. Causal conjunctions in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

'Because' is by far the most prevalent causal conjunction in the corpus. In all of its occurrences, it appears in the middle of a sentence to connect two clauses, it never occurs at the beginning of a sentence. E.g. *in some cases the level of ferritin can be misleading because we may get high level in the cases of liver damage*. Conversely, the conjunction 'since' is placed at the beginning

of a sentence in most cases (in 14 out of 20 occurrences), e.g. *Since the virus was only recently discovered, there is no certain data*…. In total, there were 86 occurrences of the word 'as' in the corpus; however, as a conjunction expressing causality it appeared 15-times, e.g. *as the most important cause of iron deficiency is bleeding*.

In the Medical Web Corpus, 'because' is the most frequent causative conjunction with 20,171 occurrences. Although 'since' has 12,977 occurrences, in most cases it is used as a preposition of time.

Adverbs were less commonly used to express causality in the corpus than conjunctions. Table 5 depicts causative adverbs found in the corpus.

adverb	no. of occurrences
1. therefore	10
2. thus	6
3. hence	3

Table 5. Adverbs expressing causation in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

There were 10 occurrences of 'therefore', and it was at the beginning of a sentence only in one instance. 'Therefore' is the 24th most frequent adverb in the corpus. Besides 'therefore', only two adverbs, 'thus' and 'hence' were found in the corpus to express causation. In the Medical Web Corpus, these same three adverbs are the most common causative adverbs; however, their order of frequency is different. 'Thus' is the most common causative adverb (n= 12,562), followed by 'therefore' with 9,591 occurrences and 'hence' (n= 2,240).

Table 6 shows the causative prepositional phrases detected in the corpus.

prepositional phrase	no. of occurrences
1. because of	16
2. due to	13
3. as a result of	6

4. as a consequence of	1

Table 6. Prepositional phrases expressing causation in the UPMS Learner Corpus of Summaries (compiled by the author based on data elicited from Sketch Engine)

Prepositional phrases expressing causation also appeared rarely in the corpus, compared to causative verbs, nouns and conjunctions. 'Because of' was the most frequent prepositional phrase (n=16). It appeared at the beginning of a sentence three times. 'Due to' was the second most common prepositional phrase to express a causative function. Out of its 13 occurrences, it appeared only once at the beginning of a sentence. These same prepositional phrases were also the most prevalent in the Medical Web Corpus; however, 'due to' (n= 15,510) was more than twice as common as 'because of' (n= 7,201).

Discussion

The function of causation plays a crucial role in biomedical texts, as it contributes to understanding the relationship between diseases, symptoms, processes, disease pathology and treatment. The high frequency of causative verbs and nouns in the Medical Web Corpus also bolsters the important role assigned to causation in biomedical texts. Therefore, it is justifiable that teaching causation receive an important role in the instruction of Medical English. This study intended to explore how Hungarian medical students who study English as a foreign language employ causation when they describe diseases. The results of the study revealed that study participants resort to a wide range of linguistic units expressing causation: causative verbs, nouns, adverbs, conjunctions and prepositional phrases. The study has shown that verbs were predominantly used to express causation, which is in line with Drury et al. (2022).

A comparison between the analysis of the UPMS Corpus of Summaries and the Medical Web Corpus reveals both similarities and differences. While Hungarian medical students primarily resorted to verbs to express causal relations, in the Medical Web Corpus causative verbs were only slightly more common than causative nouns. 'Cause' was the most common causative verb in both corpora; however, the frequency of 'cause' is much more salient in the learner corpus. The verb 'cause' is the most unambiguous way of connecting a cause and an effect (Haase 2015), which accounts for its high frequency in both corpora and its prevalence in the learner corpus. Causative verbs were much more commonly used in the active than in the passive voice in both corpora. Haase (2015) investigated the distribution of the verb 'cause' in the active and passive voice and found that the active was slightly more common in popular bioscience texts.

Conversely, in his study the passive was found to be much more common in academic bioscience texts and in both popular and academic physics texts. Learners' preference for the active voice suggests that they put emphasis on the agent (i.e. the doer) when describing causative relations. An explanation for the learners' predilection for the active voice may be that in disease descriptions concrete causers of diseases, symptoms and complications can be identified. Nevertheless, the learners' preference for the active despite a tendency to prefer the passive in scientific and academic writing, calls for further investigation.

A conspicuous difference between the usage of causative verbs in the two corpora is that in the Medical Web Corpus the use of 'produce' as a causative verb is prevalent, it is the second most common causative verb following 'cause'. Conversely, it is used to denote causation in only one instance in the learner corpus. Thus, the verb 'produce' is highly underrepresented in this corpus. 'Result' was markedly the most common causative noun in the Medical Web Corpus. On the other hand, it occurred in the learner corpus only 12-times, indicating that it is underrepresented in the learner corpus. 'Cause' was the predominant causative noun in the learner corpus. In both corpora, conjunctions, adverbs and prepositional phrases were less frequently employed to express causation compared to verbs and nouns. The same causative adverbs and prepositional phrases can be observed in both corpora; however, their order is different.

Conclusions

The finding that several similar features can be observed between the learners' use of causation and that of writers in the Medical Web Corpus suggests that the learners' use of causation is appropriate at an intermediate level. However, the overrepresentation of 'cause', both as a verb and as a 'noun', indicates that other causative verbs (e.g. 'produce') and nouns should receive more emphasis when teaching causation. An investigation of nouns has shown that study participants are not always aware of the semantic distinctions of causative words, such as the difference between 'cause' and 'reason'. In addition, in some cases they do not use prepositions correctly, especially 'for' and 'of' following causative nouns. These results suggest that both the semantic features and the syntactic patterns of linguistic units expressing causation should be refined in teaching cause and effect relationships.

Works Cited

Akkasi, Abbas and Moens, Mari-Franchine. "Causal Relationship Extraction from Biomedical Text Using Deep Neural Models: A Comprehensive Survey." *Journal of Biomedical Informatics 119* (2021): 1-14.

Altenberg, Bengt. "Causal linking in spoken and written English." *Studia Linguistica 38*(1984): 20–69.

Blanco, Eduardo, Castell, Nuria and Moldovan, Dan. "Causal Relation Extraction." In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA), 2008.

de Spinoza, Benedict. *Ethics*. London: Penguin, 1996.

Degand, Liesbeth. "Causal connectives or causal prepositions? discursive constraints." *Journal of Pragmatics 32(6).* (2000): 687–707.

Drury, Brett, Goncalo Oliveira, Hugo, de Andrade Lopes, Alneu. "A Survey of the Extraction and Applications of Causal Relations." *Natural Language Engineering 28* (2022): 361-400.

Evans, Alfred S. "Causation and Disease: The Henle-Koch's Postulates Revisited." *A Yale Journal of Biology and Medicine 49* (1976): 175-195.

Girju, Roxana. "Automatic detection of causal relations for question answering." In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*. (2003): 76–83.

Gopalan, Sindhuja, Devi, S. L. "Cause and Effect Extraction from Biomedical Corpus." *Computacion y Sistemas 21* (2017): 749–757.

Granger Sylviane. Learner corpora. In Lüdeling, A., Kytö, M. (eds.). *Corpus Linguistics. An International Handbook.* Volume 1. Berlin & New York: Walter de Gruyter, 2008. 259-275.

Haase, Christoph. "Causal Chaining and the Active/Passive Ratio in Science Texts." *Discourse and Interaction* 8/2 (2015): 21-33.

Holland, Paul W. "Statistics and Causal Inference." *Journal of American Statistical Association 81* (1986): 945-960.

Lorenz Gunter R. "Learning to cohere: causal links in native vs. non-native argumentative writing." In *Pragmatics and Beyond New Series.* (1999): 55–76.

Mihăilă, Claudiu, Ohta, Tomoko, Pyysalo, Sampo, Ananiadou, Sophia. "BioCause: Annotating and Analysing Causality in the Biomedical Domain." *BMC Bioinformat 14* (2013): 2.

Sharma, Raksha, Palshikar, Girish, Pawar, Sachin. "An Unsupervised Approach for Cause-Effect Relation Extraction from Biomedical

Text." In: *International Conference on Applications of Natural Language to Information Systems, Springer* (2018): 419-427.

Sinclair, John. Corpus and Text - Basic Principles. In Wynne, Martin. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005. 197-213.

Thagard, Paul. *How Scientists Explain Disease*. Princetown: Princetown University Press, 1999.

Trimble, Louis. *English for Science and Technology: A Discourse Approach*. Cambridge: CUP, 1985.

Vendler, Zeno. "Causal relations." *The Journal of Philosophy 21* (1967): 704–713.