Natalia Levshina*

Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions

Abstract: The present paper discusses connectivity and proximity maps of causative constructions and combines them with different types of typological data. In the first case study, I show how one can create a connectivity map based on a parallel corpus. This allows us to solve many problems, such as incomplete descriptions, inconsistent terminology and the problem of determining the semantic nodes. The second part focuses on proximity maps based on Multidimensional Scaling and compares the most important semantic distinctions, which are inferred from a parallel corpus of film subtitles and from grammar descriptions. The results suggest that corpus-based maps of tokens are more sensitive to cultural and genre-related differences in the prominence of specific causation scenarios than maps based on constructional types, which are described in reference grammars. The grammar-based maps also reveal a less clear structure, which can be due to incomplete semantic descriptions in grammars. Therefore, each approach has its shortcomings, which researchers need to be aware of.

Keywords: causation, Multidimensional Scaling, graph theory, cluster analysis, parallel corpus

1 Aims of the paper

Semantic maps serve several different purposes. First, they can represent co-verbalization patterns in a particular semantic or pragmatic domain and predict the synchronic and diachronic variation of linguistic expressions. Another purpose is to reveal cross-linguistically predominant semantic or pragmatic distinctions and clusters of functions that are frequently expressed by the same constructions in different languages. In addition, semantic maps provide a convenient tool for comparison of semantically related constructions in different languages.

*Corresponding author: Natalia Levshina,** Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, e-mail: Natalia.Levshina@mpi.nl

There exist different types of semantic maps. The two most popular ones are called connectivity maps and proximity maps. Connectivity maps (van der Auwera 2013) represent related semantic or pragmatic functions (uses, meanings, etc.) as nodes of a graph, which are linked in the most parsimonious way. These maps can represent coverbalization patterns in different languages in a clear and elegant way. They also help to predict which meanings and functions can be co-expressed by a particular construction, and which combinations are unlikely.

Proximity maps are usually based on parallel corpora. They show individual corpus instances of specific constructions. On such maps, the proximities between individual corpus instances represent the chances of these contexts being expressed by similar linguistic forms (Wälchli and Cysouw 2012; Levshina 2015). Proximity maps are useful for identification of common dimensions of semantic variation and clusters of related meanings. They can also be used to compare semantically related constructions in one language and across languages.

The main aim of this study is to discuss new hybrid approaches, crossing the boundaries between the corpus-linguistic and traditional methods in language comparison. What will happen if we use corpus data to build connectivity maps, and grammar descriptions to create proximity maps? We will see that these hybrid approaches can be fruitful and revealing. The paper contains two case studies. In the first one, I introduce a new method of using corpus data in order to create connectivity maps. In the second one, I show how one can build proximity maps with the help of Multidimensional Scaling not only based on parallel corpora, but also based on grammar descriptions, which is seldom done. I will compare the results of the two approaches and discuss what kind of theoretical and practical issues and caveats we can face while creating and interpreting these maps.

This paper will focus on causative constructions. Causation is a semantic domain which includes diverse situations where one participant of an event expressed by a clause brings about a physical or mental change in another participant (or at least has an opportunity of doing so). Consider an example in (1):

(1)   *Harry Potter made the chair levitate.*

In this sentence, Harry Potter is the Causer, and the chair is the Causee. The causing event is something that Harry did (e. g., using a levitation charm), and the caused event is that the chair levitated as a result of Harry's actions.

Causation can be expressed by various causative constructions – e. g., lexical (*break*), morphological (e. g., *solidi-fy*) or periphrastic/analytic (e. g., *make X disappear*). This semantic domain has been extensively studied in linguistic typology

and in functional and cognitive linguistics (e. g., Comrie 1981; Kemmer and Verhagen 1994; Song 1996; Dixon 2000; Talmy 2000; Shibatani and Pardeshi 2002). Numerous semantic distinctions and categories have been discussed. For example, causation can be direct or indirect, forceful or natural, intentional or accidental, factitive or permissive.

Despite that, we still do not have a detailed and empirically supported connectivity map of causative constructions. Given the massive number of studies, this may seem surprising. However, if one considers the reasons below, one can easily understand why creating a semantic map of causatives is by no means a trivial task.

One problem is the divergent terminology used by the authors of grammars and research articles. One of the problematic cases is direct and indirect causation (cf. Shibatani and Pardeshi 2002). These labels have been used to refer to numerous closely related distinctions. One of them is spatiotemporal integration of causing and caused events (Comrie 1981; Haiman 1983; cf. Fodor 1970). Another is whether the Causer is the main source of energy for the caused event or some other entity or force (e. g., Kemmer and Verhagen 1994). Although in most cases these distinctions would overlap, this is not always so. For example, a patient with a paralyzed hand can request the nurse to help him unlock his mobile phone and get some important information. According to the spatiotemporal integration criterion, this would be an instance of direct causation by the patient because his thumb would touch the screen, while the main source of energy criterion would say that the patient is not the main source of energy, and therefore the causation is indirect.

At the same time, other labels have been used to refer to very similar distinctions, such as distant vs. contact causation (Nedjalkov and Silnitsky 1973) or manipulative vs. directive causation (Shibatani and Pardeshi 2002). This inconsistency makes the cross-linguistic comparison of causatives a difficult task.

Another problem is the focus of most grammar descriptions on the most typical factitive implicative causation (i. e., making something happen or someone do something) and scarce information about the other types: permissive, accidental curative, assistive, directive, non-implicative, etc. Many grammars describe the default causative construction simply as a means of increasing the valency of a verb, without referring to its semantic functions. It remains unclear whether this construction expresses all causation types or only a subset of them.

Yet another, and less obvious, problem is that causative constructions are usually described in terms of salient distinctive features. For example, one can often find in a grammar a statement that construction A expresses direct causation, while construction B expresses indirect causation. Leaving aside the terminological issues, the problem here is that direct and indirect causation represent a pair

of contrasting features, which in language use are combined with other features. For example, indirect causation can be permissive, i. e., causation by letting X do something (e. g., *I let her read the novel*) or factitive (e. g., *I had him read the novel*), while direct causation may be forceful (e. g., *I pried the door open*) or not (e. g., *I opened the door with a key*). Both direct and indirect causation can be intentional (e. g., *I opened the door a little in order to hear the conversation* [direct, intentional] or *I had the butler open the door* [indirect, intentional]) or not intentional (e. g., *Children can open the door accidentally* [direct, not intentional] or *I caused the door to open by conjuring a wrong charm* [indirect, not intentional]), and so on. The number of distinctions mentioned in the descripttions of causatives in different languages is very high (see Section 3.1).

This focus on one or two salient dimensions represents a problem for the creation of semantic maps because their nodes are semantic functions, or uses, such as Recipient or Direction for dative constructions (Haspelmath 2003) or Specific Unknown and Specific Known for indefinite pronouns (Haspelmath 1997). In contrast, distinctive binary (or tertiary, etc.) features of causation are constantly intersected by other distinctions, which often matters for the choice of the specific causatives, as in the examples above. This makes the study of co-verbalization of causatives tricky.

To solve this problem, we could come up with an etic grid which contains all possible combinations of distinctive features (cf. Evans 2010), described and illustrated in a very transparent way. Since the grammars do not describe the semantic range of causatives in sufficient detail (see Section 3.1), one would need to fill in a questionnaire or perform an experiment. That is not feasible. Not only is the number of relevant semantic dimensions very high, but also the number of cells in the etic grid increases in geometric progression. For example, if we take 7 binary distinctions, the total number of combinations to describe would be $7^2 = 49$. The most systematic attempt has been made by Bellingham et al. (2020). They use experimental stimuli that vary along several dimensions: presence or absence of mediation (Causee, Instrument or none), type of the Causer (human + intentional, human + unintentional or natural force), type of the Affectee (intentional, human + reflexive, human + physical impact or inanimate), force dynamics (causation vs. letting), and some other distinctions. But even their impressive list of stimuli does not include all combinations and dimensions.[1] Obviously, it is a practical challenge. Moreover, even if we could obtain these data, the seman-

---

1 For example, there is no combination of human intentional Causer, human intentional Affectee and resulting activity (e. g., *The general ordered the soldiers to run*). Mental caused events (e. g., remind X of Y) are also missing, because they are difficult to show in video stimuli.

tic map would be too complicated and multidimensional to be useful as a visual tool.

In addition, due to the strong correlations between the features (Levshina 2016), some of the configurations will be unlikely to occur in the real world. As an illustration, consider forceful indirect unintentional physical causation, where the Causer exerts extra effort in order to affect the Causee indirectly and without an intention of doing so. Fans of the Harry Potter films might remember the scene where Harry gets so angry at his aunt Marge, who made very rude remarks about his parents, that he loses control and cast on her an inflating charm, making her blow up like a balloon and float about the room. Language users or experts would have problems with such items in a questionnaire because these situations are beyond their daily language experience. Also, it is unlikely that such nodes would be very useful for descriptive purposes.

The present paper offers a practical, data-driven solution to these problems. The data are taken from a parallel corpus. This allows us to avoid the problem of missing descriptions and conflicting labels. The semantic functions, which serve as the nodes, are not pre-determined in advance, but emerge as clusters from individual causation events. The clustering enables us to avoid the proliferation of nodes. The semantic map presented here is preliminary and based on a limited number of observations: it serves as a proof of concept. The method is discussed in Section 2.

As already mentioned, one purpose of semantic maps is the identification of cross-linguistically salient semantic distinctions and clusters of semantic functions or exemplars. There has been some research in this direction with the help of proximity maps. For instance, Levshina (2015) shows that a crucial distinction made by European analytic causatives is the contrast between factitive causation (making) and permissive causation (letting). The present paper focuses on all types of causatives constructions (analytic, lexical and morphological) in typologically diverse languages. Moreover, it compares two sources of data: descriptions in grammars and research articles, and parallel corpora. The question is, do the most important distinctions depend on the type of data, and, if yes, how can we explain these differences? This case study is presented in Section 3.

The statistical analyses discussed below were performed with R, a free statistical software and programming environment (R Core Team 2018), and add-on packages *cluster* (Maechler et al. 2018), *igraph* (Csardi and Nepusz 2006) and *smacof* (de Leeuw and Mair 2009). The datasets and R code can be found in the supplementary materials, available in the online version of this article.

2 Case study 1: A proof-of-concept for a corpus-based connectivity map of causatives

2.1 Data from a parallel corpus

This section outlines a data-driven approach to creating connectivity maps based on a parallel corpus. I used data from a corpus of film subtitles. All subtitles were collected from the website opensubtitles.org. Most of the subtitles were aligned with the subtitles in English with the help of software subalign (Tiedemann 2012). The advantage of this type of data is its closeness to spontaneous informal discourse (Levshina 2017). Psycholinguistic research has demonstrated that film subtitles provide lexical frequency norms, which are known to outperform other text sources in predicting different behavioural measures, such as lexical decision latencies (Keuleers et al. 2010; but see a more critical view in Baayen et al. 2016). The main disadvantage is the use of translations, rather than original texts, in most languages, which may create translationese effects. This is a well-known problem in parallel corpus research (cf. Cysouw and Wälchli 2007). Another limitation is the restricted number of characters and lines in subtitles. However, there is no evidence that this represents a problem for causative constructions. From the author's personal experience, the frequencies of major formal and semantic types of causatives in subtitles and spontaneous conversations are comparable.

For this pilot study, I took 18 causative situations from the English subtitles of the film *Avatar* (2009), more exactly, its first half. The reason for this choice was that this film has a rich inventory of diverse causation types, which involve human minds and bodies, animals and plants, natural and supernatural forces, weapons, military troops and corporations. Moreover, *Avatar* has many subtitle translations in different languages, thanks to its numerous fans. The list of situations, which were selected randomly, is given in Table 1. Many instances of direct factitive causation expressed by lexical causatives (e. g., *kill*, *cut*, *break*) were ignored because this is by far the most common and cross-linguistically uniform type of causation. Some situations were discarded because they had no causative constructions in the translations.

Next, I manually found the constructions that express these causation events in 22 languages, including the English version. The majority are Indo-European languages: Germanic (English, Danish, Dutch, German, Swedish), Romance (French, Portuguese, Romanian, Spanish) and Slavic (Bulgarian, Czech, Polish and Russian). The non-Indo-European languages are Uralic (Estonian, Finnish and Hungarian), Hebrew, Indonesian, Mandarin Chinese, Thai, Turkish and Vietnamese.

Table 1: List of causation events and their short labels.

English sentence	Short label
Let's go special case, do not make me wait for you.	You_make_me_wait
As head of security, it is my job to keep you alive.	I_keep_you_alive
Plus it'll help keep (you sane).	It_HELPS_keep_you_sane
(Plus it'll help) keep you sane.	It_helps_KEEP_you_sane
Just relax and let your mind go blank.	Let_your_mind_go_blank
You're wiggling your toes!	You_wiggle_toes
This low gravity will make you soft.	Gravity_makes_you_soft
(They could fix me up, if I rotated back.)	They_make_me_pretty
Make me pretty again.	
It reminds me every day what's waiting out there.	It_reminds_me_of_X
Just keep your mouth shut (and let Norm do the talking).	Keep_mouth_shut
(Just keep your mouth shut and) let Norm do the talking.	Let_him_do_talking
Relax marine, you're making me nervous.	You_make_me_nervous
Norm, you've contaminated the sample with your saliva.	You_contaminate_sample
Why not let them just kill my ass?	Let_them_kill_my_ass
Who gets them to move?	Who_gets_them_to_move
So just find me a carrot that will get them to move.	Carrot_gets_them_to_move
You may tell her what to do, inside.	You_tell_her_what_to_do
I'm not about to let Selfridge and Quaritch micromanage this thing.	I_let_SMB_micromanage_thing

The translations found in the subtitles were analysed and coded as different types of causative constructions. All lexical causatives were treated as one constructional type. Morphological causatives were distinguished depending on the suffix, e. g., Indonesian has causative verbs with the suffixes *-kan* and *-i*, for instance *meng-hamil-kan* and *meng-hamil-i*, which both mean 'make pregnant' (Sneddon 1996: 97). Such examples were treated as different constructions. Analytic causatives were distinguished by the verb that expresses the abstract causing event. For example, French causatives *faire* + Infinitive and *laisser* + Infinitive were treated as separate constructions. A non-trivial question is how fine-grained the classification should be. For example, one may wonder whether to treat *make* + *Verb* and *make* + *Adjective* as one or two constructions. In the European languages, where adjectives and verbs are easy to distinguish, I treated such constructions as separate types, while in the South-East Asian languages, where this distinction is not obvious, they were considered as one type. Also, finite and non-finite predicates specifying the caused event were treated differently in those languages where one could make this distinction. Versions of causative verbs with and without prefixes were treated as one construction, e. g., Indone-

Table 2: Matrix of constructions and causation events: a fragment.

Language	Construction	You_make_me_wait	I_keep_you_alive	It_HELPS_keep_you_sane
ENG	make_Vinf	1	0	0
ENG	keep_Adj	0	1	0
ENG	help_Vinf	0	0	1
ENG	let_Vinf	0	0	0
ENG	make_Adj	0	0	0
ENG	Lex	0	0	0
ENG	get_toVinf	0	0	0
ENG	tell_toVinf	0	0	0
RUS	zastavljat_Vinf	1	0	0
RUS	pomogat_Vinf	0	0	1
RUS	Lex	0	0	0
RUS	derzhat_Adj	0	0	0

sian (*mem*)*buat* 'make' + Predicate. The marking on the Causee and word order patterns were disregarded. Of course, the problem of deciding what constitutes a separate construction is by no means limited to corpus data. When relying on grammars, we "outsource" the decision to the author, which does not make the issue any less problematic.

The result of this analysis was a matrix of language-specific constructions (rows) and causation events (columns). A fragment is shown in Table 2. In this table, '1' means that this construction was used to express the causation event in the corpus, whereas '0' means that this construction was not used to express this causation event.

2.2 Identification of nodes

In order to identify more abstract semantic functions, which can serve as the nodes of a semantic map, the causation events were clustered according to how they are expressed in the languages. First, a distance matrix of the situations was created based on the similarity measures. The values in the table were treated as binary asymmetric, which means the following. If two causation events were expressed with the same construction in a given language, the similarity score was increased by 1. If not, nothing was added. The sum similarity between the events was then divided by the total number of comparisons between the causation events where at least one value was non-zero. This proportion is also known as Jaccard similarity coefficient J. The similarity scores were transformed into distances by subtracting the former from 1: $D = 1 - J$. After that, a standard agglomerative cluster analysis was performed using the average method of ag-
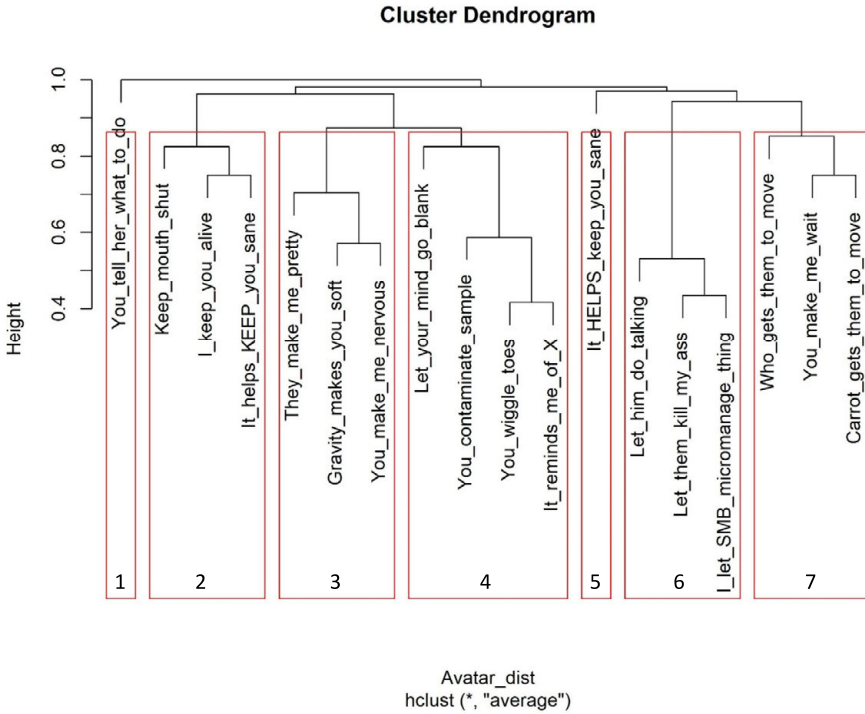
Cluster Dendrogram

Avatar_dist
hclust (*, "average")

Figure 1: Cluster analysis of 18 causation events.

gregation. Figure 1 displays the clustering tree. It is advisable to try out different clustering methods in order to see if the results converge.

The figure displays seven clusters, delimited by red rectangles. Why was this number chosen? One can determine the optimal number of clusters using so-called average silhouette widths. The higher the silhouette width, the better the clustering solution. A good solution means that the members of the clusters are maximally close to one another and at the same time maximally distant from the members of the other clusters. Figure 2 shows the silhouette widths for the number of clusters from 2 to 10. The plot suggests that the optimal number of clusters is 7.

The clusters in Figure 1 can be interpreted as follows, moving from left to right:

- Cluster 1. TELL: mandative, or directive, non-implicative causation (You tell her what to do), which is only represented by one context. Non-implicative means that it is not clear whether the caused event took place or not (Karttunen 1971). Note that this type of causation does not overlap formally with any other causation events. All other examples represent implicative causation.
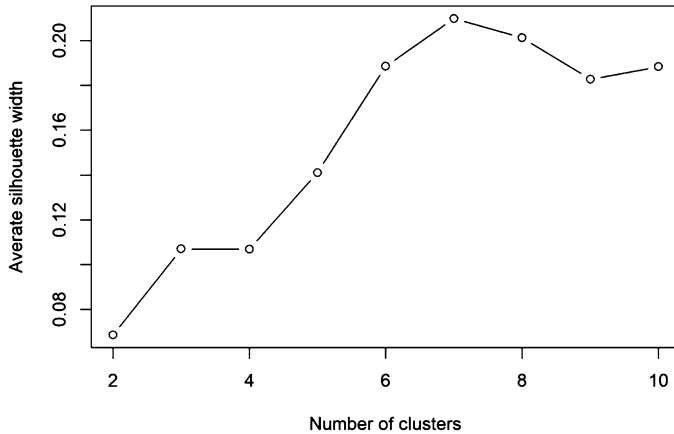
Figure 2: Average silhouette widths for different number of clusters.

– Cluster 2. KEEP: keeping the animate or inanimate Causee in a particular state, by a human or some circumstances.
– Cluster 3. MAKE_X: the animate or inanimate Causer intentionally or accidentally makes the Causee change its state. The Causee in the examples is animate, but affected by the Causer.
– Cluster 4. NO_CONTROL_CE: somebody or something brings about the change in the Causee, which has no control over the process.
– Cluster 5. HELP: only one example, where something helps to keep someone in a certain state.
– Cluster 6. LET: causation of non-interference (Talmy's 2000 non-impingement) of the Causer in the action performed by the Causee.
– Cluster 7. CONTROL_CE: indirect factitive (i. e., not permissive) causation in which the Causee has control over the caused event. The Causer can be animate or inanimate, acting intentionally or not intentionally.

Let us now consider more closely the neighbouring clusters 3 and 4 with MAKE_X and NO_CONTROL_CE. The formal difference between them is that MAKE_X is expressed in many languages as a verb of making followed by an adjective. As for NO_CONTROL_CE, most languages express these types with lexical causatives (with some exceptions, such as the English original *Let your mind go blank*, which corresponds to a lexical causative in some other languages). This explains why these two semantic types do not form one big cluster. This formal difference can be explained by the differences in usage frequencies (Haspelmath 2008). Lexical causatives are usually more frequent than analytic ones. Language users exhibit
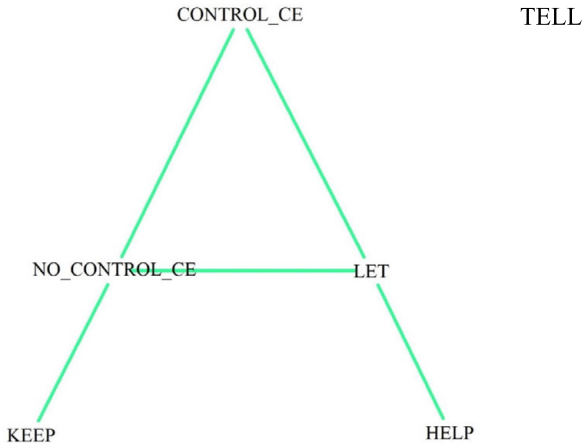
Figure 3: Semantic map of causatives. CE stands for Causee.

communicatively efficient behaviour, choosing shorter words to express frequent causation events, and leaving more cumbersome analytic structures for less frequent events (Levshina 2018). Since frequency differences between the nodes are usually not relevant for connectivity maps, these two functions can be merged. In what follows, the label NO_CONTROL_CE will also cover the cluster MAKE_X. Whether to merge clusters or not depends on the focus of one's study.

2.3 Automatic creation of the semantic map

After the clusters were identified, the original construction-event matrix was recoded. The columns were now the six big clusters described above (where MAKE_X is treated as an instance of NO_CONTROL_CE). If a construction was used to express at least one event from a cluster, it received value 1. Based on this matrix, we will build the semantic map. For this purpose, I used the software created by Regier et al. (2013).[2] Their Python script, which builds parsimonious semantic maps from a matrix of linguistic forms and semantic functions, is based on an efficient algorithm for the social network inference problem proposed by Angluin et al. (2010). The result is shown in Figure 3. The position of TELL, which has no co-expression links, was chosen arbitrarily.

---

2 The Python code can be found at http://lclab.berkeley.edu/regier/semantic-maps/ (22 April 2021).
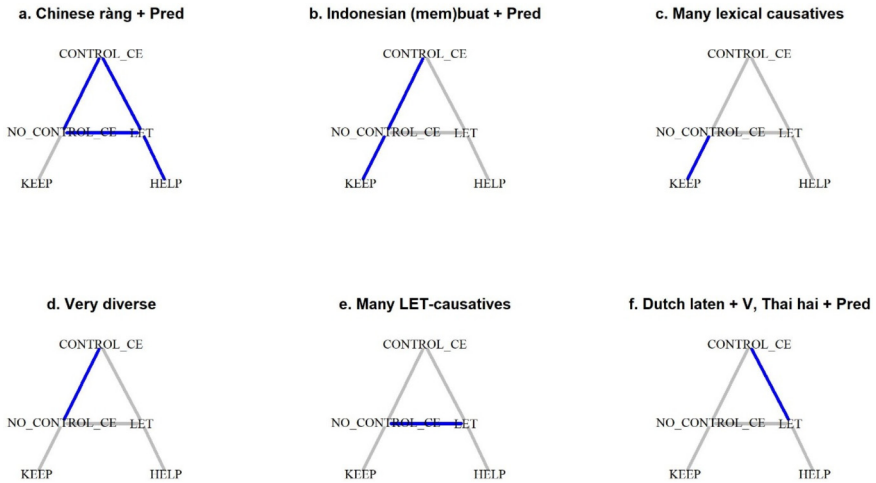
Figure 4: Semantic configurations of different language-specific constructions.

Figure 4 demonstrates how the language-specific constructions map onto the resulting graph. Note that the lone node TELL is not displayed. The largest semantic area belongs to the Chinese analytic causative *ràng* + Predicate. As shown in Figure 4a, it has all of the semantic functions with the exception of KEEP. It is followed by the Indonesian analytic causative (*mem*)*buat* + Predicate, which was attested in the clusters KEEP, NO_CONTROL_CE and CONTROL_CAUSEE (see Figure 4b). KEEP and NO_CONTROL_CE are found in many lexical causatives: Bulgarian, Danish, German, Hebrew, Indonesian and Romanian, as shown in Figure 4c. The combination of NO_CONTROL_CE is found in numerous constructions: Finnish and Turkish morphological causatives, Portuguese *fazer* + Infinitive, Vietnamese analytic causative with *làm*, and lexical causatives in Finnish, Hungarian, Russian and Thai (Figure 4d). NO_CONTROL_CE and LET are co-expressed by the English *let* + Infinitive, Indonesian *biarkan* + Predicate, Portuguese *deixar* + Infinitive and Finnish *antaa* + Infinitive (Figure 4e). LET and CONTROL_CE co-occur in the Dutch *laten* + Infinitive and Thai *hai* + Predicate (Figure 4f). The remaining constructions have only been observed in one function and therefore do not contribute anything to the map.

The map allows us to make some predictions. In particular, a causative construction that expresses KEEP is unlikely to express LET if it does not express direct factive causation NO_CONTROL_CE. Similarly, HELP cannot be co-expressed with the factive causation functions if LET is not expressed, as well. As for TELL, it is likely to be merged with the rest of the map when more

data are added. The reason is the fact that some languages develop an implicative causative from originally non-implicative constructions with directive semantics (i. e., order/ask/tell X to do Y). An example is Polish *kazać*, which means 'say, tell', but which can be also used implicatively, as in (2):

(2)   *Kazał-Ø   na siebie   długo czekać*[3]
      say.PST-3SG.M on self.ACC long   wait
      'He made (others) wait for himself for a long time.'

It is most likely that TELL will be merged with CONTROL_CE first because the addressee of a request or order can decide whether to comply with the Causer's demands or not.

   This case study has demonstrated a solution for creating connectivity maps based on corpora. It can help us overcome the challenges that arise when studying the semantics of causative constructions (see Section 1). Obviously, this is only a first step towards building a data-driven connectivity map of causative constructions. We need more different sources of data, more non-European languages and more causative situations to cluster. In particular, we need original texts in other languages and other genres, in order to control for translationese and the space restrictions associated with subtitles.

3 Case study 2: Comparing proximity maps based on grammar descriptions and parallel corpora

3.1 A proximity map based on grammar descriptions

3.1.1 Typological data from grammar descriptions

Causation is a standard item on the to-do list of grammarians who plan to describe a language. However, there are a few serious problems with the existing descriptions, which were discussed in Section 1. This type of data is not entirely useless for semantic maps, however. We can focus on the cross-linguistically salient distinctions, as done in previous token-based proximity maps (see Section 2). Which types of causation are usually expressed by different constructions, and which by similar ones in languages of the world? We will use a popular dimensionality reduction technique, Multidimensional Scaling (MDS) (e. g., Croft and Poole 2008;

---

3 https://sjp.pwn.pl/szukaj/kaza%C4%87.html (22 April 2021).

Wälchli and Cysouw 2012). This procedure will help us visualize the distinctions that are most typical in languages of the world.

More than 200 reference grammars were analysed for this case study. From those, descriptions of 142 causative constructions were obtained which satisfied two criteria. First, there was a semantic description. Second, each of the languoids belonged to a unique genus, according to the classification given in the online World Atlas of Language Structures (Dryer and Haspelmath 2013). These constructions represent 62 languoids and therefore 62 genera.

The semantic descriptions had very diverse semantic categories, which were classified into several larger superordinate categories, as shown in Table 3. This generalization was a necessary step. Without it, we would not be able to compare the constructions cross-linguistically. Note that some of the features have their opposites in the descriptions (e. g., DIRECT vs. INDIRECT), whereas the others do not (e. g., PORTATIVE or ASSISTIVE).

Table 3: Semantic distinctions found in grammars and their superordinate categories (in capitals).

DIRECT	INDIRECT
direct, contact, manipulative, non-mediated, causer as the main source of energy or controller responsible on the caused event, focus on the effect on the Causee, integrated cause and effect, physical contact between the Causer and Causee, the Causee is affected, direct personal involvement of the Causer; describes what happens to the Causee, not what he/she does; affected Causee	indirect, distant, roundabout or 'indefinite' causation, the events are weakly or not (necessarily) integrated spatiotemporally; 'X arranged the matter so that Y happened', interpersonal causation
INTENTIONAL	NOT INTENTIONAL
intentional or purposeful causation, volitional or agentive instigator/Causer	accidental, unintentional
ASSISTIVE	
assistive; the causer is in the position of caring for, chaperoning, or helping the Causee	
CURATIVE	
curative, have something done by someone	
NO CONTROLLING CAUSEE	CONTROLLING CAUSEE
non-controlling, non-volitional or passive Causee, the Causee's volition or choice is downgraded, the Causee is affected, has no causal agency	the Causee is active and controlling; the Causee performs some action, has a degree of autonomy, interpersonal causation
IMPLICATIVE	NON-IMPLICATIVE

Table 3 (continued)

COMITATIVE	
comitative, sociative or involved causation; the causer performs or experiences the action or state he/she/it initiates, participating to some degree in the joint activity	
NOT FORCEFUL	**FORCEFUL**
non-forceful, mild, effortless or neutral causation, cause something without resistance	forceful causation, forcing, causation by force
COMMUNICATION	
causation with the help of (verbal) communication: directive, mandative (command, ordering, compelling, persuading), rogative (asking), invitation	
FACTITIVE	**PERMISSIVE**
factitive or coercive causation, making	permissive causation, letting, can express 'allow'
CAUSED STATE	**CAUSED ACTION**
causation of a state	
PORTATIVE	
portative	
PUTTING	
causation as putting in a certain position	
TOTALLY AFFECTED CAUSEE	
total affectedness of the object	
INTENSIVE	
intensive	
MOVING CAUSEE	
moving Causee or liquid	
DOUBLE CAUSATION	
three-participants causation	

In addition to that, the following features were added to the data based on conceptual considerations:
– Permissive causation was also coded as indirect;
– Mandative/directive and curative causation were also coded as indirect and with a controlling Causee;
– Forceful causation was treated as factitive (not permissive);
– Portative causation (carry X to some place) was treated as comitative (i. e., involving the Causer who also performs the caused event): by carrying or bringing something to a certain location, the Causer also moves there;
– Causation with a controlling Causee is indirect (with the exception of assistive causation);
– Double causation (e. g., X causing Y to do X) is treated as indirect.

3.1.2 Semantic map based on MDS

In order to obtain stable results, all rare features that occur in less than 5 constructions were removed.[4] Also, all constructions in which only one feature is observed, were disregarded because they do not contribute anything to the distances between the semantic features. The resulting matrix with data constituted 98 rows (language-specific constructions) and 18 columns (semantic features).

The presence of a feature in the semantic range of a construction was coded as 1, and its absence was coded as 0. These values were treated differently, i. e., the data were binary asymmetric (see an explanation of this term in Section 2.2). The reason is that the absence of a feature (0) in a construction's description may be due to missing information or due to a genuine absence of this function in the semantic repertoire of the construction. In most cases, we cannot tell. So, when two features have 0 in one construction and 0 in another one, we cannot treat that as a sign of correlation between the features. Similarity is only established when both features have 1 in both constructions.

Next, I performed MDS based on stress minimization using majorization (de Leeuw and Mair 2009), which produces low stress – a standard measure of goodness of fit of an MDS solution. The default ratio scaling method was used. I tested several solutions with a different number of dimensions and chose the solution with three dimensions as the optimal one. The stress of the 3D solution was 0.21. Adding the 4th dimension did not help to identify new patterns, and did not result in a large decrease in stress.

The existing analytical tools for MDS allow us to investigate which of the points are poorly represented by the map. A stress plot suggests that Forceful and Curative are the features whose distances to the other points are the least reliably fitted by the MDS. However, none of the features is responsible for more than 8 % of the total amount of stress, which means that none of the features has a very poor fit.

The resulting MDS is shown in Figure 5, from two different perspectives. The upper plot displays Dimensions 1 and 2, whereas the bottom plot shows Dimensions 1 and 3.

Looking closely at the maps, we can see that the dimensions are quite difficult to interpret. The first dimension (horizontal) stretches from comitative/assistive/curative causation on the left to permissive and indirect causation on the

---

4 The excluded features were PORTATIVE, PUTTING, DOUBLE CAUSATION, TOTALLY AFFECTED CAUSEE, MOVING CAUSEE. Of course, there is a chance that these features were simply not discussed systematically in grammars, so it is safer to treat them as descriptive *rara*, rather than as typological ones.
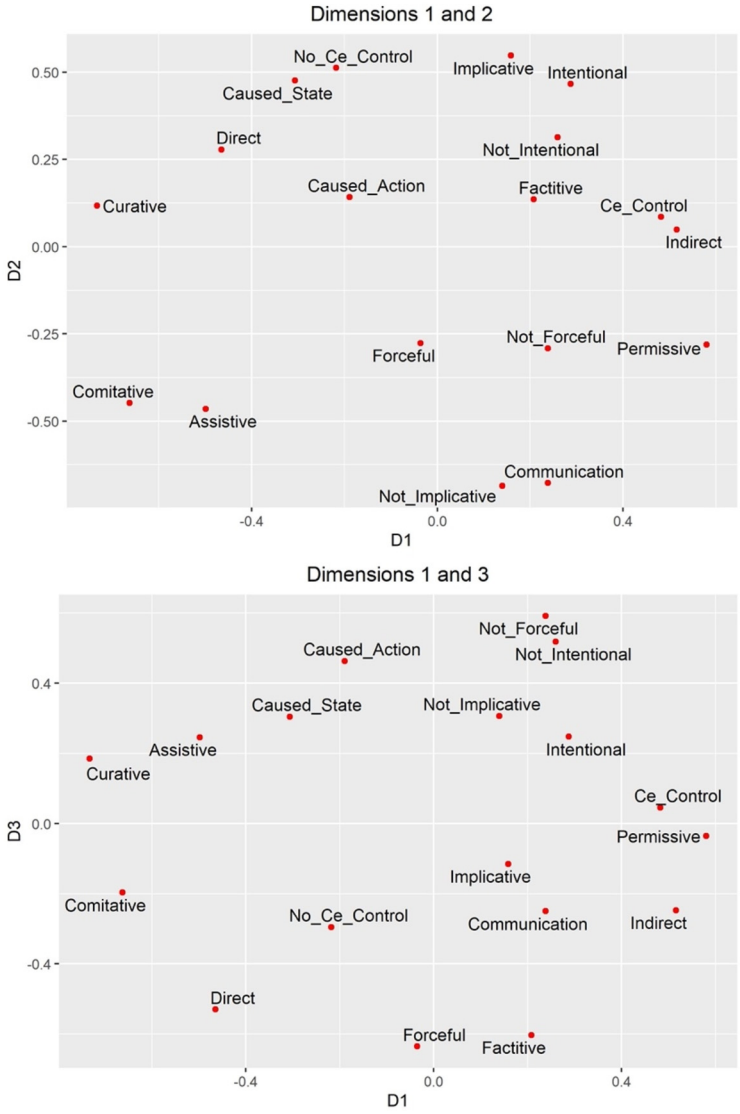
Figure 5: A MDS solution Top: dimensions 1 and 2; bottom: dimensions 1 and 3.

extreme right. It also correlates with direct (left) vs. indirect causation (right). It is difficult to find a common denominator here. The second dimension shown in the upper plot contrasts implicative intentional causation without Causee's control at the top with non-implicative causation at the bottom, which often goes with cau-

sation by communication (i. e., directive, mandative, etc. causation). Finally, the third dimension places forceful causation at the bottom and non-forceful at the very top. Forceful causation is normally factitive, but non-forceful is usually associated with communication, assistance, curative causation and caused actions where the Causee is an agent.

The lack of clear structure in the MDS solution supports our previous impression that causation, at least, the way it is described in grammars, does not have obvious prototypes, which could be easily selected as nodes on a connectivity map. It suggests that there are many different dimensions and distinctions that cross-cut each other. To what extent this is an artifact of descriptions, which often omit correlated features, or a real property of very diverse causative systems in the world's languages is a question that we cannot answer now. In order to be able to answer it in the future, we need to develop and consistently use comparative concepts for description of causation semantics. Table 3 can serve as a starting point. We should also pay systematic attention to peripheral semantic distinctions.

3.2 Proximity map based on a parallel corpus

3.2.1 Parallel corpus data

This section presents an MDS map based on data from parallel corpora. For this case study, I again used a corpus of film subtitles. There were five films of different genres:

- *Avatar* (action, adventure, fantasy, 2009)
- *Black Swan* (drama, thriller, 2010)
- *Frozen* (animation, adventure, comedy, 2013)
- *Noah* (action, adventure, drama, 2014)
- *Twilight* (drama, fantasy, romance, 2008)

As in the case study discussed in Section 2, the subtitles were collected from the website opensubtitles.org and aligned with the English version. In addition to English, nine other languages from different genera were investigated: Finnish, French, Hebrew, Indonesian, Mandarin Chinese, Russian, Thai, Turkish and Vietnamese.

As a starting point, 182 contexts were found in the English version (without sampling). These contexts represent diverse causative events (factitive and permissive, implicative and non-implicative, assistive, curative, etc.). The constructions expressing these events were identified in each of the languages. See more information about the coding in Section 2.1.

Table 4: Data from the parallel corpus: a fragment.

Film	Context	RUS	FIN	FRA	THA	VIE	TUR
Avatar	*They can **fix** the spine, if you got the money.*	Lex	Lex	Lex	Lex	Lex	Morph
Avatar	*Let's go special case, do not **make** me **wait** for you.*	zastavl-jat'_V	Morph	*faire*_V	*hai*_Pred	*để*_Pred	NA
Avatar	*…that will **stop** your heart in one minute.*	Lex	Morph	NA	Lex	NA	Morph
Avatar	*As head of security, it is my job to **keep** you **alive**.*	NA	*pitää*_V	NA	NA	NA	Pred_*tut-mak*
Avatar	*It's nothing like an old school safety brief to **put** your mind **at ease**.*	Lex	Lex	*met-tre*_PP	Lex	NA	NA
Avatar	*Me and Norm here, are to **drive** these remotely controlled bodies called avatars.*	Lex	Lex	Lex	Lex	Lex	Lex
Avatar	*but you can **take** them **out** tomorrow.*	Lex	Lex	NA	Lex	Lex	NA
Avatar	*Don't shoot, you'll **piss** him **off**!*	Lex	Morph	Lex	*tamhai*_Pred	*lam*_Pred	Morph

Consider an example. One of the contexts in *Avatar* was *Don't shoot, you'll piss him off*. The causative event was 'SMB pisses SMB off'. In English, the construction was encoded as a lexical causative. In French, the translation was *Vous allez l'énerver*, which also contains a lexical causative. The Turkish translation was *Onu kızdıracaksın*, where the verb *kızdır-* is a morphological causative derived from *kız-* 'lose one's temper'. The Vietnamese version contained an analytic causative with the auxiliary *làm* 'make'

(3)   Cậu sẽ làm nó nổi   điên đó
        you will make him become mad there
        'You'll cause him to get mad.'

The result of this time-consuming procedure, in which I relied on numerous reference grammars and online translation tools, was a matrix with the causation events as rows and the languages as columns. The cells contained the construction types. See a fragment of the table in Table 4. If a non-causative periphrastic expression was used or the translation was incomplete or erroneous (in rare cases), the label NA (not available) was added.

Table 5: Adjusted R^2 of the linear regression on the coordinates of each dimension.

Semantic distinction	Dimension 1	Dimension 2	Dimension 3
Intentional or Not	−0.004	0.071	0.032
Forceful or Not	0.011	0.004	−0.005
Comitative or Not	0.003	−0.004	0.001
Direct or Not	0.191	−0.004	0.000
Causee Control or No Control	0.203	−0.001	0.015
Factitive or Permissive	**0.520**	**0.158**	0.045
Caused Event or State	0.290	−0.003	0.012
Implicative or Not	0.021	−0.006	0.037
Assistive or Not	0.019	0.022	**0.147**
Curative or Not	0.005	−0.005	−0.005
Directive or Not	0.021	−0.006	0.038

3.2.2 MDS map

Based on the matrix described in the previous section, I created a matrix of distances between the causative events. The distances were computed as follows. If two causative events had the same constructions within one language, 1 was added to their similarity score. The sum similarity score was then divided by the total number of comparisons without missing data. Finally, the proportion was subtracted from one.

As in the previous case, solutions with the number of dimensions 2, 3 and 4 were tested. The solution with 3 dimensions was optimal, with the stress value of 0.16. Adding the fourth dimension would only reduce the stress by 0.03, which is a very small improvement.

Interpreting an MDS map with corpus tokens is always a difficult task. In order to make the interpretation more objective, I coded the causation situations for the same features as in the previous case study. The annotation involved a visual inspection of the relevant scenes in the films.

Next, I used linear regression analysis in order to identify which of the semantic variables are most strongly correlated with the coordinates of the causative events on the three dimensions of the MDS model. The measure was adjusted R^2, which is the standard way of representing the relative strength of relationship between two variables in linear regression. The procedure has already been used by Levshina (2011) for similar purposes. The coefficients are shown in Table 5.

The coefficients reveal that the first dimension is associated most strongly with the distinction between factitive and permissive causation (i. e., making vs. letting), followed by the features related to (in)directness of causation and the role
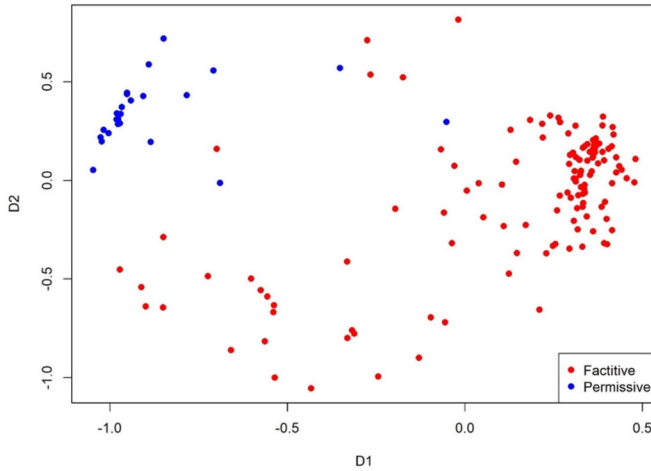
Figure 6: MDS based on corpus data: dimensions 1 and 2.

of the Causee. The second dimension is also correlated with factitive vs. permissive causation, and to some extent with the distinction between intentional and non-intentional Causer. Figure 6 displays the 1st and 2nd dimensions of the MDS model, which show that the permissive events form a separate cluster in the top left corner. Note that the dense cluster on the right corresponds to direct factitive causation, usually expressed by lexical causatives (and sometimes morphological ones). Finally, the third dimension reveals a contrast between assistive (i. e., helping) and non-assistive causation due to the cluster of situations related to helping at the bottom of Figure 7, which shows the second and the third dimension.

Thus, different types of data yield different salient semantic dimensions. In particular, the distinction between factitive and permissive causation is very prominent in the corpus – a contrast we do not clearly observe in the grammar data. This may be explained by two reasons. First, according to Wierzbicka (2006: Sect. 6.2.3), letting is an important category in Anglo-Saxon culture because it is associated with non-interference, non-imposition and personal freedom. In addition to actual letting as absence or cessation of impingement, e. g., *I let him enter/escape/stay*, one can also find instances of interpersonal letting in the data, where the verb *let* is used to regulate human interaction. These uses can be classified into several types:

- *let* of permission, i. e., not preventing the Causee from doing what he or she intends to do, e. g., *I'll let you tag along* (Frozen);
- *let* of shared information, e. g., *Let me know if that's juicy enough for you* (Black Swan);
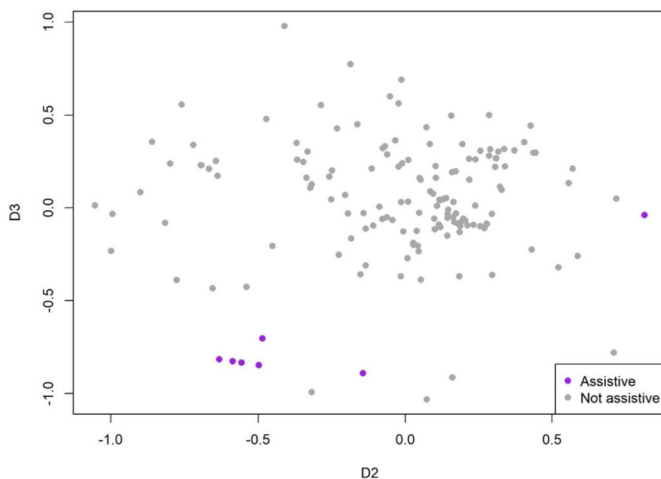
Figure 7: MDS based on corpus data: dimensions 2 and 3.

–   *let* of tolerance, e. g., *So I'll let you finish. Bye* (Black Swan);
–   *let* of cooperative dialogue, e. g., *Let me make this very important announce-
    ment* (Black Swan);
–   *let* of cooperative interaction, e. g., *Let me take you to Noah* (Noah).

According to Wierzbicka (2006: 187), these uses "acknowledge the rights of the ad-
dressee as an autonomous and free individual, on equal footing with the speaker."

This prominence of letting in the English data may have an impact on its fre-
quency in the other languages, which often use periphrastic causatives with verbs
of permission to translate *let* in contexts of non-interference (in particular, per-
mission and tolerance). For instance, *I'll let you finish* is translated by a construc-
tion with the verb *dat'* 'give' in Russian, *antaa* 'let' in Finnish, *để* 'let' in Viet-
namese, etc. The example of *let* of cooperative interaction *Let me take you to Noah*
is translated in some languages by a verb of letting, e. g., French *Laissez-moi vous
emmener voir Noé*, which may not be the best choice. Such translationese effects
may increase the influence of the making vs. letting distinction on the geometry of
the MDS solution. At the same time, many of the other instances of cooperative let-
ting are not expressed by such constructions. For instance, the above-mentioned
example of *let* of cooperative dialogue *Let me make this very important announce-
ment* is translated into French as *J'ai une annonce très importante à faire!* 'I've got
a very important announcement to make!'. The request for shared information *Let
me know* is translated mostly by lexical verbs of saying or telling. These choices
are idiomatic.

Another important reason for the absence of a prominent contrast between factitive and permissive causation in the typological data may be the predominant focus of the grammarians on factitive causation as the default causation type. This can be explained by stereotypes of what constitutes causation proper, as well as by the higher level of grammaticalization of more frequent factitive causation in comparison with relatively infrequent permissive causation. As a result, the grammarians may not regard permissive causation as a part of the grammatical system, leaving it for the lexicon.

Unlike in the typological data, we do not find a clear contrast between forceful and non-forceful causation in the corpus – probably because this type of causation is rare in the selected films and in our everyday life, thanks to different machines, gadgets and social institutions. There can be a cultural difference between different languages, as far as the prominence of physical effort is concerned. In post-industrial societies, physical violence and hard physical work have become less common than previously due to cultural and technological improvements.

From this case study we learn that the important dimensions of variation and semantic distinctions strongly depend on the type of data. We should be extremely careful when making general claims based on data from one source. The idea that both parallel corpora and descriptive grammars only represent doculects (a blend of *documented* and *lects*), rather than languages as such (Cysouw and Good 2013), is particularly important here. It may well be possible that there is no common map of causation for different types of texts and genres, so the goal of inferring common dimensions of variation (see Section 1) may not be attainable, after all. Given the diversity of semantic distinctions and the cultural malleability of event conceptualizations, this is a plausible scenario.

4 Conclusions and implications

The aim of this paper was to discuss different combinations of data and methods in creation of semantic maps. In the first case study, I proposed a method that can help linguists to create semantic maps based on corpus data, rather than grammatical descriptions. It was argued that the creation of a connectivity map of causatives based on grammatical descriptions is extremely difficult. The reasons are the focus of grammars on the most salient distinctive features, the terminological confusion and the incomplete descriptions of causatives. The paper has shown how one can solve these problems with the help of a data-driven approach, using data from a parallel corpus and inferring the nodes as clusters of individual causative events. We hope that future research will enrich this preliminary map

with new functions and links between them. We will need more diverse parallel corpora with different source languages and genres for that purpose.

In the second case study, the focus was on another important goal of semantic maps, that is, identification of salient semantic contrasts in a semantic domain. The comparison of MDS based on typological and parallel corpus data (Section 3) has revealed that the salient distinctions vary depending on the data. This may have to do with the sparse descriptions in the grammars, in particular, the focus on the highly grammaticalized default causative forms, which usually express factitive implicative causation. Also, there is a substantial bias towards lexical causatives in the subtitles, whereas they are often ignored in the grammars, which focus usually on morphological causatives.

The differences may also be due to cultural factors. According to Wierzbicka (2006), Anglo-Saxon culture emphasizes the importance of non-interference, non-imposition and negative freedom, which can be regarded as negative politeness. This explains why English has a variety of constructions with interpersonal *let* (Wierzbicka 2006: Sect. 6.2.3). The high frequency of permissive causation in the English version of the film subtitles may explain why the contrast between factitive and permissive causation is very strong in the corpus data, and weak in the grammar data. Moreover, forceful causation may be less important in the highly industrialized cultures than in other societies, which could explain why there are few tokens of this type in the corpus data and therefore no evidence that this distinction is important, unlike in the typological data. Another reason might be that violence is normally shown on the screen, rather than described verbally in the films. In general, one would expect corpus-based semantic maps with usage tokens to be more prone to predominant cultural scenarios and genre effects than maps based on grammars and constructional types. Therefore, it would be worthwhile to look at different pivot languages, films from other countries and other genres of text in the future. On the other hand, as already mentioned, grammars are also biased towards the prototypical causation types (direct, intentional, factitive, non-assistive, etc.).

From all this follows that one should be extremely careful when trying to interpret a semantic map as representing some universal conceptual space. A universal space presupposes the same dimensions. However, we have seen that they may differ substantially depending on the type of data. These results suggest that at least for some conceptual domains, the goal of identifying the most important dimensions of variation for languages, rather than for doculects, may be unattainable. This does not mean that such maps are useless. They can be a convenient tool for comparison of semantics of different constructions. An example is identification of grammaticalizaton clines in Romance and Germanic causative constructions in Levshina (2015).

Finally, the results of this study have implications for the description of causatives in reference grammars. The authors should be encouraged to focus on more dimensions of causation than simply (in)directness and related distinctions, such as intentional vs. non-intentional, implicative vs. non-implicative causation, assistive vs. non-assistive causation. This also means that they will need to include more diverse constructions, going beyond the most typical causatives (usually morphological ones). In addition, it is necessary to create comparative concepts representing causative semantics. The concepts displayed in Table 3, for example, could serve as a starting point. This would help to avoid the existing terminological inconsistency in the description of causatives.

References

Angluin, Dana, James Aspnes & Lev Reyzin. 2010. Inferring social networks from outbreaks. In Markus Hutter, Frank Stephan, Vladimir Vovk & Thomas Zeugmann (eds.), *Algorithmic learning theory 2010* (Lecture Notes in Computer Science 6331), 104–118. Berlin: Springer.

Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* 30(11). 1174–1220. https://doi.org/10.1080/02687038.2016.1147767.

Bellingham, Erika, Stephanie Evers, Kazuhiro Kawachi, Alice Mitchell, Sang-Hee Park, Anastasia Stepanova & Jürgen Bohnemeyer. 2020. Exploring the representation of causality across languages: Integrating production, comprehension and conceptualization perspectives. In Elitzur Bar-Asher Siegal & Nora Boneh (eds.), *Perspectives on causation. Selected papers from the Jerusalem 2017 workshop*, 75–119. Cham: Springer. https://doi.org/10.1007/978-3-030-34308-8_3.

Comrie, Bernard. 1981. *Language universals and linguistic typology: Syntax and morphology*. Chicago, IL: The University of Chicago Press.

Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics* 34. 1–37.

Csardi, Gabor & Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695. http://igraph.org (22 April 2021).

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion 'language'. *Language Documentation and Conservation* 7. 331–359.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung (STUF)* 60(2). 95–99.

de Leeuw, Jan & Patrick Mair. 2009. Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software* 31(3). 1–30. https://doi.org/10.18637/jss.v031.i03.

Dixon, R.M.W. 2000. A typology of causatives: Form, syntax and meaning. In R.M.W. Dixon & Alexandra Y. Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 30–83. Cambridge: Cambridge University Press.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info (22 April 2021).

Evans, Nicholas 2010. Semantic typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 504–533. Oxford: Oxford University Press.

Fodor, Jerry. 1970. Three reasons for not deriving 'kill' from 'cause to die'. *Linguistic Inquiry* 1(4). 429–438.

Haiman, John 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language: Cognitive and functional approaches to language structure*, Vol. 2, 211–242. Mahwah, NJ: Lawrence Erlbaum Associates.

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.

Karttunen, Larry. 1971. Implicative verbs. *Language* 47(2). 340–358.

Kemmer, Suzanne & Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5(2). 115–156.

Keuleers, Emmanuel, Marc Brysbaert & Boris New. 2010. SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods* 42. 643–650.

Levshina, Natalia. 2011. *Doe wat je niet laten kan [Do what you cannot let]: A usage-based analysis of Dutch causative constructions*. Leuven: University of Leuven dissertation.

Levshina, Natalia. 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520.

Levshina, Natalia. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2). 507–542.

Levshina, Natalia. 2017. Online film subtitles as a corpus: An n-gram approach. *Corpora* 12(3). 311–338.

Levshina, Natalia. 2018. *Towards a theory of communicative efficiency in human languages*. Leipzig: Leipzig University habilitation thesis.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2018. *cluster: Cluster analysis basics and extensions*. R package version 2.0.7-1.

Nedjalkov, Vladimir P. & Georgiy G. Silnitsky. 1973. The typology of morphological and lexical causatives. In Ferenc Kiefer (ed.), *Trends in Soviet theoretical linguistics*, 1–32. Dordrecht: Reidel.

R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (22 April 2021).

Regier, Terry, Naveen Khetarpal & Asifa Majid. 2013. Inferring semantic maps. *Linguistic Typology* 17(1). 89–105.

Shibatani, Masayoshi & Prashant Pardeshi. 2002. The causative continuum. In Masayoshi Shibatani (ed.), *The grammar of causation and interpersonal manipulation*, 85–126. Amsterdam: Benjamins.

Sneddon, James N. 1996. *Indonesian: A comprehensive grammar*. New York: Routledge.

Song, Jae Jung. 1996. *Causatives and causation: A universal-typological perspective*. London & New York: Addison Wesley Longman.

Talmy, Leonard. 2000. *Toward a ccognitive semantics*. Cambridge, MA: MIT Press.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (eds.), *Proceedings of the Eighth International Conference*

on Language Resources and Evaluation (LREC-2012), 2214–2218. Istanbul: European
    Language Resources Association (ELRA).
van der Auwera, Johan. 2013. Semantic maps, for synchronic and diachronic typology. In
    Anna G. Ramat, Caterina Mauri & Pietra Molinelli (eds.), *Synchrony and Diachrony:
    A dynamic interface*, 153–176. Amsterdam: Benjamins.
Wälchli, Bernhard, and Michael Cysouw. 2012. Lexical typology through similarity semantics:
    Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.
Wierzbicka, Anna. 2006. *English: Meaning and culture*. Oxford: Oxford University Press.