Linguistic repercussions of COVID-19 : A corpus study on four languages

Carter, Emmanuel

2022-12-05

Carter , E , Onysko , A , Winter-Froemel , E , Zenner , E , Gisle , A , Hilberink-Schulpen , B , Nederstigt , U , Peterson , E & van Meurs , F 2022 , ' Linguistic repercussions of COVID-19 : A corpus study on four languages ' , Open Linguistics , vol. 8 , no. 1 , pp. 751-766 . https://doi.org/10.1515/opli-2022-

http://hdl.handle.net/10138/352254
https://doi.org/10.1515/opli-2022-0222

Research Article

Emmanuel Cartier, Alexander Onysko*, Esme Winter-Froemel, Eline Zenner, Gisle Andersen, Béryl Hilberink-Schulpen, Ulrike Nederstigt, Elizabeth Peterson, and Frank van Meurs

Linguistic repercussions of COVID-19: A corpus study on four languages

https://doi.org/10.1515/opli-2022-0222
received October 25, 2021; accepted October 17, 2022

Abstract: The global reach of the COVID-19 pandemic and the ensuing localized policy reactions provides a case to uncover how a global crisis translates into linguistic discourse. Based on the *JSI Timestamped Web Corpora* that are automatically POS-tagged and accessible via SketchEngine, this study compares French, German, Dutch, and English. After identifying the main names used to denote the virus and its disease, we extracted a total of 1,697 associated terms (according to *logDice* values) retrieved from news media data from January through October 2020. These associated words were then organized into categories describing the properties of the virus and the disease, their spatio-temporal features and their cause–effect dependencies. Analyzing the output cross-linguistically and across the first 10 months of the pandemic, a fairly stable semantic discourse space is found within and across each of the four languages, with an overall clear preference for visual and biomedical features as associated terms, though significant diatopic and diachronic shifts in the discourse space are also attested.

Keywords: COVID-19, French, English, Dutch, German, collocates, semantic categorization, pandemic discourse

1 Introduction

In December 2019, the World Health Organization (WHO) received various signals concerning a cluster of cases of pneumonia with an unknown etiology in Wuhan, China.[1] WHO was asked by several health authorities from around the world to provide additional information. By January 2020, it was clear that the reported cases of pneumonia were caused by a novel virus with evidence of human-to-human transmission. By the end of January 2020, the first cases of "coronavirus" disease, named after the crown-like

1 This rudimentary COVID pandemic timeline is based on the WHO interactive timeline (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#! , consulted 21 April 2021) and on the overview presented in Aslam et al. (2020, 3).

*** Corresponding author: Alexander Onysko,** Department of English, University of Klagenfurt, Klagenfurt 9020, Austria, e-mail: Alexander.Onysko@aau.at
Emmanuel Cartier: Département de linguistique, Université Sorbonne 13 Paris Nord, Paris, France
Esme Winter-Froemel: Neuphilologisches Institut/Romanistik, University of Würzburg, Würzburg, Germany
Eline Zenner: Quantitative Lexicology and Variational Linguistics, KU Leuven, Leuven, Belgium
Gisle Andersen: Department of Professional and Intercultural Communication, Norges Handelshøyskole, Bergen, Norway
Béryl Hilberink-Schulpen, Ulrike Nederstigt, Frank van Meurs: Department of Language and Communication, Radboud University, Nijmegen, The Netherlands
Elizabeth Peterson: Department of Languages, University of Helsinki, Finland

thorns on the surface of the virus, were attested outside China, in the United States and Europe, serving as a clear warning sign of a potential global spread of the virus. By the end of March 2020, the virus was dispersed across much of the globe, with at least 730,000 cases of patients with confirmed COVID-19[2] and over 36,000 reported deaths. From that time onward to the time of writing (October 2021), the COVID-19 pandemic dominated national and global policies, with differing responses to control the spread of the virus and limit the pressure on healthcare as much as possible. These responses, which evolved across time and varied considerably between countries and regions, included campaigns emphasizing physical distance, face masks, and hand hygiene, as well as lockdowns that closed down most of the public lives and confined people to their homes. Further frequent consequences were extensive testing of those with and without COVID-19 symptoms and a range of second-order measures to accommodate the economic and social impact of the primary measures.

Language has taken on a key role in dealing with this worldwide phenomenon that has profoundly affected everyday life. The importance of language is not restricted to the terminological choices needed to label the new virus and the disease it causes (see, e.g., terminological efforts such as the IATE database and its definition of the term coronavirus[3]). Language is also there to conceptualize, describe, and hence socially and discursively negotiate the existence of the virus, its effects, and the measures and reactions related to it. Given the global spread of the virus but the often localized policy reactions to the consequences of the pandemic, it is pivotal to understand better how the pandemic translates into linguistic discourse and to verify what cross-linguistic similarities and differences can be attested in the linguistic treatment of the novel coronavirus.

This article addresses that question by pursuing the following three objectives: (1) identifying the terms used to denote the virus and the disease,[4] (2) identifying the main words associated with these terms, and (3) comparing the results for (1) and (2) across the first 10 months of the pandemic for four languages.

The article will first give an overview of previous research on language during the COVID-19 pandemic in order to identify gaps and motivate the research questions that will be addressed in this article (Section 2). The methodological approach of the study will be outlined in Section 3, and the results will be discussed in Section 4, focusing on the major names for the virus and the disease and their associated words in the four languages under scrutiny.

2 Language in the COVID-19 pandemic: previous research and open questions

There is an emerging abundance of research on language use relating to the COVID-19 pandemic in its first year. Overall, two main strands of research can be identified: (1) studies trying to keep track of the lexical impact of the pandemic; (2) studies interpreting linguistic data as a manifestation of the social and cultural reception and consequences of the pandemic.

Research trying to keep track of how SARS-CoV-2 has affected the word stocks of given languages can itself be subdivided into lexicographical and lexicological studies on the one hand and natural language processing approaches on the other hand. First, efforts to keep track of coronavirus-related neologisms have been made by linguists, lexicographical institutes, news media agencies, language standardization bodies, and lexicologists. Examples include the glossaries of coronavirus terminology published by *Time*, *Oxford English Dictionary* and *Institut für Deutsche Sprache*. The new words kept in these glossaries are

2 This term was coined to label the disease caused by the coronavirus, with 19 referring to 2019 as the year in which the first COVID case was identified.

3 https://iate.europa.eu/entry/result/3588006 for coronavirus, consulted 21 April 2021.

4 Due to the close connection of the virus (cause) and the resulting disease (effect), the names for both the virus and the disease are important in naming practices of the COVID-19 pandemic. In the study, the names for the virus and the disease will be treated individually, while they are often used synonymously in everyday language use to refer to the COVID-19 pandemic.

analyzed by lexicologists who uncover word formation processes and types of neology, typically through qualitative analysis (see, e.g., Balnat 2020 for French and German, Pietrini 2021 and Sgroi 2020 for Italian, Ladilova 2020, Rodríguez Abella 2021 for Spanish, Belhaj 2020 for French words in Moroccan newspapers, Bowker 2020 comparing COVID-19-related terms in Canadian and European French, and Roig-Marín 2021 for English). In some instances, lexicologists focus on one particular COVID-related term. For example, Thiéry-Riboulot (2020) describes diachronic patterns in the semantics of the word *confinement* and Pietrini (2020) focuses on the Italian word *distanza* 'distance'.

At the same time, researchers from the field of natural language processing started to create inventories of coronavirus-related words to facilitate the analysis of coronavirus terminology and discourse. The Coronavirus Corpus released in May 2020 provides a large, constantly updated corpus of coronavirus-related data from online newspapers and magazines in 20 different English-speaking countries (Davies 2021). Leaman and Lu (2020) have created a comprehensive dictionary of unique terms used in scientific literature to refer to SARS-CoV-2 and COVID-19. Ma et al. (2021) have adopted a broader approach that not only gathers names for the virus and the related disease but that also includes COVID-19 terminology for ten different related categories (e.g., clinical manifestation, epidemic prevention, and control), covering a total of 464 concepts for 724 related Chinese and 887 related English terms. Lew and Kosem's contribution to Tan et al. (2020) presents a tool that allows researchers to identify COVID neologisms in a timestamped corpus. Most other linguistic tools created by and for researchers concerning language related to the COVID-19 pandemic rely on Twitter data and often adopt a coarse-grained crosslinguistic perspective. Abuld-Mageed et al. (2020) released Mega-Cov, a billion-scale geo-located Twitter data set including data from over 65 languages (see also Lopez et al. 2020, Chen et al. 2020a).

The databases of tweets and keywords provided by natural language processing serve as input to the second main strand of research we can identify, namely studies analyzing linguistic manifestations of the social impact and reception of the COVID-19 pandemic (e.g., Chen et al. 2020b working with the database of Chen et al. 2020a). A methodological divide characterizes research adopting more quantitative and more qualitative perspectives although in both cases, the goal is usually to conduct a content analysis of COVID-related messages, to trace potentially stigmatizing social stereotypes occurring in COVID discourse, and to pinpoint the emotional load of COVID language, often with attention to the small-scale diachronic trends to be witnessed in the first months of the pandemic.

In the quantitative realm, content analysis is conducted automatically, with researchers relying on (combinations of) raw frequency of tweets (Singh et al. 2020), raw frequency of particular strings in tweets (Abuld-Mageed et al. 2020 on the most frequent hashtags), topic models (e.g., Kurten and Beullens 2021), and keyword identification (Makhachashvili and Bilyk 2020, using SketchEngine on a French news media corpus, see also Spina 2020). Wicke and Bolognesi (2020) verify in which of the automatically identified COVID topics WAR metaphors are most frequently attested, compared to the occurrence of alternative source domains for COVID (STORM, MONSTER, and TSUNAMI, also compare Semino 2021). In a preprint paper, Solovejetu and Gatherer (2020) expand on the typically rather limited diachronic span of COVID content analysis. The authors use corpus linguistic tools to look for signs of previous respiratory flues in historical texts. As a second pursuit, quantitative approaches aim to uncover patterns of stigmatization and social stereotypes in COVID discourse. For example, Hu et al. (2020) and Budwhani and Sun (2020) analyze stigmatizing references to the virus through location or origin (*Wuhan*, *China*, *Chinese*). Finally, natural language processing approaches have paid attention to evolutions in the sentiment and emotions of COVID discourse. While Kurten and Beullens (2021) integrate sentiment analysis in their more general content analysis of a Twitter corpus, Aslam et al. (2020) use automatic classification of emotions for a dataset of over 140,000 COVID newspaper headlines and show high negative polarity in the headlines.

A range of qualitative approaches also sheds more light on language use as indicative of the social impact during the first year of the COVID-19 pandemic. Attention to metaphors is found in the study by Craig (2020) and Semino (2021), both again critically stressing the pervasiveness of WAR metaphors in discourse on the pandemic. Nossem (2020) studies how names of the virus can be seen as doing linguistic rebordering; the disease becomes framed as something foreign and hence as "a threat to the nation from the outside" (Nossem 2020, 77). Black (2020), Chun (2020), and Du (2020) describe markers of stigma, stereotyping, and racial profiling in COVID discourse. Parvin et al. (2020) and Katermina and Yachenko (2020)

carry out a qualitative keyword approach to analyze the content of COVID discourse in mass media corpora. Cougnon and de Viron (2020) explicitly compare findings for a corpus of lay tweets versus tweets from political bodies versus tweets from mass and news media organizations. A Critical Discourse Analysis approach is found in MIRCo (2020), where a research group reflects on entries they made in a quarantine discourse diary. Finally, Zhang and Li (2020)'s special issue on *Multilingua* contains several papers on multilingual and translanguaging practices in light of COVID-19 (e.g., Piller et al. 2020), describing how global tendencies in the pandemic and local interpretations are intertwined in social media discourse (Zhu 2020, Zhang and Zhao 2020) and also in particular linguistic practices such as poetry (Chen 2020) or Mongolian fiddle stories (Bai 2020).

To sum up, the research presented here can be classified according to three criteria. First, in terms of data, many researchers focus on social media (particularly Twitter), with some researchers working on mass media corpora. More research on large newspaper corpora reflecting international news coverage appears to be needed. Second, studies typically emphasize either inventorization or interpretation. Work in the lexicographical tradition is oriented toward a description from world to language; that is, it starts from the new virus and the concepts it has introduced and describes the way lexical building blocks are used to create new names for these concepts. Studies in the tradition of content and (Critical) Discourse Analysis, on the other hand, go the other way, from language to world; that is, they depart from the linguistic representation of the virus and the language found in its vicinity in order to uncover what it reveals about our framing of reality. What is needed though are studies combining both perspectives, i.e., objectively addressing how new realities are named and discussed, while also analyzing the impact of general semantic categories in the emerging discourses. This study aims at addressing both aspects by combining a cross-linguistic corpus study of four languages with fine-grained usage-based lexical semantic analyses. As such, we aim to draw on the benefits of large-scale corpora, yet go beyond the typically coarse-grained content classification of automatic procedures to bridge the divide between world knowledge and linguistic knowledge, integrating inventorization and interpretation. Following Parvin et al.'s (2020, 1) claim that "[d]uring all critical incidents, the media frame our understanding and create powerful forces at both individual and societal levels," the article will address the following research questions:

RQ1: What is the distribution of different competing names for the virus and the disease it causes in various languages (i.e., *virus*, *corona*, *coronavirus*, *covid*, *covid19*, *ncov*, *sars*, and *SARScov2*), considering the effect of the WHO official name (COVID-19) and the unfolding of events before and after this naming?

RQ2: What are the most common semantic categories attested in the associated words surrounding the main names for the coronavirus and the disease it causes (i.e., *virus*, *corona*, *coronavirus*, *covid*, *covid19*, *ncov*, *sars*, and *SARScov2*) in a cross-linguistic news media corpus?

RQ3: Does the distribution found in answer to RQ2 show variation when adopting a contrastive perspective, comparing patterns of associated words for Dutch, English, French, and German (3.1), when adopting a short-term diachronic perspective, comparing the evolution of associated words from January to October 2020 (3.2), and when combining the cross-linguistic and diachronic perspective (3.3)?

The next section proceeds with a discussion of the methodological building blocks of our approach.

3 Data and method

In order to address the RQs outlined above, we use a cross-linguistic corpus that allows us to chart small-scale diachronic changes in news-media discourse during the first year of the pandemic. Second, a set of names of the virus/disease ("seed terms," see below) was isolated from the corpus, and their frequency of use was determined. Next, with the help of appropriate measures, the associated words for each of these seed terms were identified. These associated words were then interpreted and classified from a semantic perspective through an iterative qualitative analysis, laying the groundwork for a quantitative analysis of cross-linguistic and diachronic evolutions.

Table 1: *JSI Timestamped Corpora* main figures (2020 subcorpora)

	Number of sources (websites)	Number of tokens
Dutch (D)	2,320	248,652,251
English (E)	45,937	8,609,763,901
French (F)	7,926	1,153,421,927
German (G)	6,588	1,090,567,908

3.1 Corpora

The selected media corpora need to be sufficiently similar in the four languages both in quantitative and in qualitative terms. The *JSI Timestamped Web Corpora* (Trampus and Blaz 2012) provide an appropriate collection of online newspaper articles automatically retrieved for about 20 languages from 2014 onward. The texts are automatically POS-tagged and stored in SketchEngine (Kilgariff et al. 2014). We use the 2020 corpora split into 10 months (January to October) for the four studied languages: Dutch, English, French, and German. The time span is due to the availability of the corpora at the time of data collection. At the same time, we assume that the first 10 months of the pandemic is a particularly relevant period as it captures how the reality of the virus and its disease became discursively encoded in a short-term diachronic span of language use. As shown in Table 1, the English corpus is notably larger than the others, and the Dutch corpus is the smallest at about a quarter of the size of the French and German ones. This bias is mitigated because the sizes of all the corpora are sufficiently large to allow for solid statistical processing. Furthermore, we applied statistical association measures which are independent of corpus size.

3.2 Seed terms, associated words, and noise

We use the notion ***seed term*** for the lexemes that are used to denote the concepts central to naming the novel virus and the disease it causes. The seed terms were identified in two ways: first, we retrieved the various scientific names for the virus from the IATE database, WHO public communications, and research papers, resulting in the following English short list: *coronavirus*, *covid-19*, *SARS-Cov-2,* and *N-Cov*. In addition, we also included non-technical names, as it has been shown that in order to name an unknown thing, people resort to existing name(s) enlisting the most (subjective) salient semantic features with the perceptive and conceptual understanding of the new thing (cf. Blank 1998). In this case, the noun *virus* together with *coronavirus* were the most frequent lexemes.[5] To sum up, the list of seed terms consists of *virus*, *corona*, *coronavirus*, *covid*, *covid19*, *ncov*, *sars*, and *SARScov2*.[6] Post-processing has been carried out to cover orthographic variations of the seed lexemes: case variations, hyphenization (*corona virus, coronavirus, coronavirus, covid-19*, and *covid19*), language-specific orthographic adaptation (*coronavirus* > *Coronavirus* (German)), common truncations (*covid-19* > *covid*, *SARS-Cov-2* > *SARS-Cov* > *SARS*), and misspellings (*SARS* > *SRAS*). In all, over 1,500,000 occurrences of the seed terms were found in the entire corpus.

For each of the seed terms, we retrieved associated words occurring significantly frequently in the neighborhood of our seed terms, set to five words to the left and five words to the right of the seed term.

5 As soon as the new virus was associated with the coronavirus family.

6 It needs to be mentioned that these names do not all exhibit a similar morpho-semantic structure: whereas *virus* is the prototypical morphologically undecomposable core name (expressing the basic level category), the compound *coronavirus* comprises a modifier and head element (*corona* and *virus*), and the others are even more complex semantically. Among those, *covid-19* is the most synthetic lexeme since it is an acronym that combines four distinct lexical units: *corona*, *virus*, *disease*, and 2019. At the same time, the morphologically complex names differ with respect to their semantic transparency: while *coronavirus* combines the internationalism *virus* with the Latin item *corona* 'crown', specifying a visual characteristic of the virus by metaphorical similarity, the elements of the acronym *covid-19* are not necessarily transparent to all speakers.

Table 2: Overview of the number of *associated words* across 10 months per language per seed term after removal of noise

	Dutch (D)		English (E)		French (F)		German (G)		Total	
	n	%	*N*	%	*n*	%	*N*	%	*N*	%
Corona	75	22.8	30	7.7	26	11.4	75	20.7	206	15.7
Coronavirus	93	28.3	78	20.1	76	33.2	90	24.9	337	25.8
Covid	11	3.3	61	15.7	34	14.8	30	8.3	136	10.4
Covid(-)19	68	20.7	84	21.6	25	10.9	54	14.9	231	17.7
Ncov	0	0.0	4	1.0	0	0.0	0	0.0	4	0.3
sars	1	0.3	23	5.9	0	0.0	12	3.3	36	2.8
sars(-)cov(-)2	1	0.3	54	13.9	0	0.0	2	0.6	57	4.4
Virus	80	24.3	54	13.9	68	29.7	99	27.3	301	23.0
Total	329	100	388	100	229	100	362	100	1308	100

This is the most commonly used lexical surrounding when looking for associated words (Evert et al. 2017). To retrieve the associated words from our corpora, we relied on an association measure. Several association measures are available for collocation extraction. Each comes with its distinctive merits, but, generally, they lead to fairly comparable results (see Pecina 2010 and Evert et al. 2017 for a review). In our study, we used *logDice*, an adaptation of the *Dice* Score (Rychlý 2008), which has the advantage of being independent of corpus size while ensuring reasonable results.[7]

To reduce the inventory of associated words to those that are most likely to effectively occur in a direct semantic relation with the seed term, we crossed this statistical measure with linguistic filtering. This filter narrowed in on grammatical relations according to which the associated word was directly dependent on the nominal seed term (adjectival or nominal modifier, adjectival, nominal, or verbal predication).[8]

The extraction of the associated words was carried out for each language, each month (January to October 2020), and each seed term. To make qualitative processing manageable, we retained only the ten most relevant associated words according to *logDice* values per month per seed term per language. If fewer than ten significantly associated words were found, all associated words were kept. At first, 1,697 associated terms were identified. This initial set was checked for noise. For example, we excluded proper names unrelated to the COVID-19 pandemic (e.g., associated words such as *Jesus*, referring to the soccer player *Jesus Corona*), seed terms that appeared as associated words (e.g., *coronavirus* as an associated term for *covid-19*), and associated words that referred to the name of another virus (*Ebola*, *flu*, *H1N1*, *flu*, and *mers*).

Table 2 shows the number of associated words per language per seed term after noise removal. Note that the numbers in the table are aggregate counts of associated words per month per seed term per language. This means that one particular word type can reoccur as an *associated word* for different seed terms and/or in different months. For instance, Dutch *besmetting* "infection" accounts for 8 of the 329 associated words identified for Dutch in Table 2. It occurs as a significantly associated word for three of our seed terms (*coronavirus*, *corona*, and *covid-19*) and in eight different months. Further, note that Table 2 does not provide any information on the number of tokens we find in the corpus for the associated words.

7 From Rychlý 2008:

$\text{Dice} = \frac{2f_{xy}}{f_x + f_y}$

$\text{logDice} = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$

where *f* is the relative frequency, *x* and *y* the seed term and the associated words. The numerator corresponds to the relative frequency of *x* and *y* multiplied by 2 (as there are two terms), and the denominator corresponds to the sum of the individual relative frequencies. When *x* and *y* only occur together, *Dice* is equal to 1. When *x* and *y* never occur together, the *Dice* score is 0. The *logDice* is an adaptation of this score to make it more readable, by first applying binary logarithm to the *Dice* score (i.e. *Dice* values below 0.5 become negative, and above 0.5 positive up to 0), to smoothen the result, and adding 14. A 0 value means that there is less than 1 occurrence of *XY* per 16,000 *X* or *Y*. Values above 7 are considered strong associations.

8 This post-processing is conducted based on results provided by the SketchEngine's WordSketch-tool.

More information can be found in Supplement A,[9] which lists the types and their token counts per language. Supplement A, for instance, reveals that the 329 associated words found for Dutch consist of 111 different word types that together occur 5,734 times in the newspaper corpus.

3.3 Semantic coding of the associated words

For our semantic classification and analysis, we looked for patterns and semantic clusters in the associated words. This allows for a conceptual analysis of similarities and differences across languages and months. To this end, a qualitative iterative procedure was followed to identify overarching semantic categories in the list of associated words. As such, the complete set of 1,308 associated words as given in Table 2 was coded in a bottom-up iterative procedure and the words were grouped according to more general semantic categories that relate to the concepts of VIRUS and DISEASE.

This iterative coding procedure led to 12 emerging semantic categories that are summarized in Table 3. A full list of associated words per language per category can be found in Supplement B. The coding for the emerging categories was done in a binary fashion: "yes" in case the associated word belongs to the semantic category and "no" in case it does not. For instance, the associated word *Wuhan* received "yes" for the category LOCATION as the semantics of the word pertains to a city in China. The word received "no" for VISUAL & BIOMEDICAL FEATURES as its meaning does not relate to this general category emerging from the data. Note that an associated word could receive "yes" for several of the categories. The German term *ausgebrochen* 'broken out', for instance, received "yes" for DIFFUSION & SPREAD (the meaning of *ausgebrochen* relates to something appearing where it was not present before), and also for the categories EVALUATION (it is a non-neutral, fairly negative description) and INTENSITY (it describes breaking out at a higher than average intensity).

Two further coding principles were adhered to. First, coding was based on the general semantic meaning of the lexical item, aggregated over individual usage contexts. Corpus concordances were resorted to if disambiguation was required, e.g., to disambiguate between the evaluative meaning of German *positiv* 'good, beneficial' (EVALUATION) and its medical meaning 'characterized by the presence of a feature' (VISUAL & BIOMEDICAL FEATURES) as expressed in the frequent collocation of "positiv getestet" ('tested positive').

Second, inter-rater reliability was achieved through a qualitative coding procedure. For each language, the data were coded by the authors of this manuscript, consisting of teams of at least two coders per language. All coders were highly proficient speakers of the language in question, and at least one coder was a native speaker of the coded language. In case of divergences, the final coding was fixed after a discussion between the coders for each of the languages and across the four languages.

4 Results

The data resulting from the coding procedure can be accessed via an interactive web interface (https://tal.lipn.univ-paris13.fr/neoveille/html/covid19_project/html/data_exploration.php), which provides an encompassing view of the database and its patterns; see Supplement C for supporting information. In this section, we restrict our attention to our research questions as stated in Section 2.

4.1 Cross-linguistic analysis of seed name distribution and evolution (RQ1)

In order to address RQ1 on the distribution of different competing names for the virus, Table 4 provides the total number of occurrences and the percentage of each of the seed terms in the corpus for each of the four languages.

9 All supplements can be accessed via https://zenodo.org/record/7339615.

Table 3: Overview of the semantic categories emerging from the data, including their definitions and examples for each

Category	Description	Examples
VISUAL & BIOMEDICAL FEATURES	The associated word describes visual, perceptual or biomedical features of the virus or disease	E *asymptomatic*, *contagious*, *respiratory* G *anhaftend* 'adhering', *hochansteckend* 'highly contagious', *infektiös* 'infectious' F *invisible* 'invisible', *mortel* 'deadly', *tueur* 'murderous' D *dodelijk* 'deadly', *levensgevaarlijk* 'life-threatening', *ziekte* 'disease'
LOCATION	The associated word indicates a geographic location or area	E *Wuhan* G *weltweit* 'world wide' F *chinois* 'Chinese' D *consultatiepunt* 'consultation check-point'
RECENCY	The associated word expresses that the item referred to (the virus/disease) is considered to be novel at the time of discourse, in the discursive context or in the conceptual world of the discourse participants	E *novel* G *neu* 'new', *neuartig* 'novel' F *nouveau* 'new' D *nieuw* 'new'
TEMPORAL SEQUENCE	The associated word establishes a diachronic relationship between several states in time	E *pre, post* G *momentan* 'current', *wöchentlich* 'weekly' F *après* 'after', *saisonnier* 'seasonal' D *weekcijfers* 'weekly figures'
DIFFUSION & SPREAD	The associated word expresses processes of spread across a community or an organism	E *contagious*, *outbreak* G *grassierend* 'rampant', *weltweit* 'world wide' F *circulant* 'circulating' D *uitbraak* 'outbreak'
MEASURES	The associated word describes medical and societal effects and measures taken in response to the disease or following the pandemic	E *contain*, *cure* G *kostenlos* 'free of charge', *negativ* 'negative' F *compatible* 'compatible', *prêt* 'loan' D *corona-app* 'corona-app', *coupe* 'hairstyle'
CAUSE & EFFECT	The associated word indexes cause and effect relations	E *effect*, *impact* G *absichtlich* 'intentional(ly)', *wirksam* 'effective' F *grâce* 'thanks to' D *impact* 'impact'
COMPARISON	The associated word establishes relations of similarity and contrast to other entities	E *flu-like* G *ähnlich* 'similar', *verwandt* 'related' F *grippal* 'influenzal', *semblable* 'similar' D *sars-achtig* 'sars-like'
EVALUATION	The associated word represents an emotionally loaded expression	E *plague*, *wretched* G *blöd* 'stupid', *heimtückisch* 'insidious' F *maudit* 'cursed', *pernicieux* 'dangerous, harmful' D *verdomd* 'cursed', *vermaledijd* 'cursed'
INTENSITY	The associated word expresses degrees of strength, impact, force, or an exceptional degree	E *moderate*, *severe* G *abgeschwächt* 'weakened', *hochansteckend* 'highly contagious' F *plein* 'entirely' D *oplaaien* 'to flare up'
(UN)CERTAINTY	The associated word relates to the degrees to which a given statement (about the virus/disease) is taken to be sure and indubitable	E *lab-confirmed* G *bestätigt* 'confirmed', *nachweislich* 'proven' F *confirmé* 'confirmed', *mystérieux* 'mysterious' D *bewijzen* 'to prove', *mysterieus* 'mysterious'
METALINGUISTIC DISCOURSE	The associated word is a name in relation to the virus/disease, or it refers to discourse about the virus/disease (e.g., in special newspaper columns or blogs about the virus/disease)	E *aka*, *Latin* G *spezial* 'special' F *latin* 'Latin', *spécial* 'special' D *dashboard* 'dashboard', *weekcijfers* 'weekly figures'

Table 4: Total number of seed terms per language in the newspaper corpus, in absolute (*n*) and relative (%) frequency

	German		English		French		Dutch	
	N	**%**	***N***	**%**	***N***	**%**	***N***	**%**
Corona	390,988	26.0	246,928	1.13	16,987	0.70	93,240	22.27
Coronavirus	464,905	30.9	7,316,596	33.53	949,745	39.04	162,820	38.89
Covid	15,320	1.02	1,062,656	4.87	161,086	6.62	3,877	0.93
Covid(-)19	286,365	19.06	9,548,732	43.76	908,236	37.3	53,894	12.8
ncov	300	0.02	10,536	0.05	595	0.02	38	0.01
sars	5,582	0.37	126,585	0.58	1,101	0.05	1,924	0.46
Sars(-)cov(-)2	55,026	3.66	147,270	0.67	19,724	0.81	2,878	0.69
Virus	284,239	18.91	3,361,834	15.41	375,181	15.42	99,952	23.88
Total	1,502,725	100	21,821,137	100	2,432,655	100	418,623	100

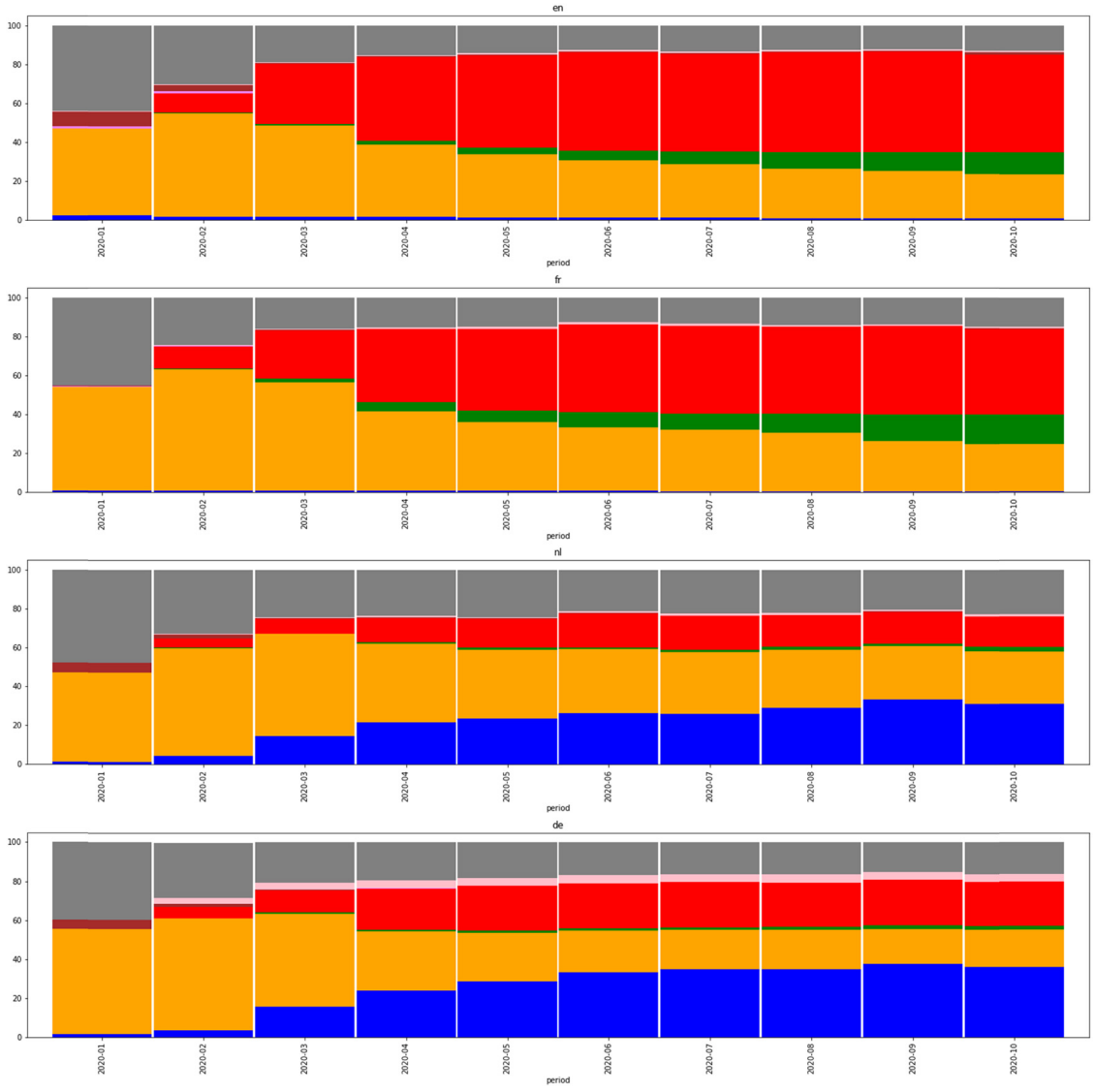

Figure 1: Diachronic view of seed terms distribution per month (from left to right: January to October), for each language (from top to bottom: English, French, Dutch, and German), expressed as a percentage (calculated from the sum of absolute frequency per language per month [color codes: blue = corona, orange = coronavirus, green = covid, red = covid-19, violet = *n*-cov, brown = sars, pink = sars-cov-2, grey = virus]).

Table 4 shows the distribution of the various seed terms used to denote the virus and the disease throughout the period for each of the languages. Figure 1 adds to that by illustrating how the distribution of the seed terms evolved across the 10 months for each language. The results demonstrate that after the virus was spotted and when it was attested in Europe, it was designated with broader terms such as *virus* and *coronavirus*. Later on, the shortening *corona* gained currency in English, Dutch, and German while in French the possible shortening competes strongly with the popular beer brand *Corona*, obstructing its use as a name for the virus/disease. When the World Health Organization devised a specific term, *covid-19*, that name was immediately adopted. The period from February to October 2020 exhibits the various ways in which this official name concurs with the already entrenched preceding words. In French and English, *covid-19* (and the shortened form *covid*) quickly diffused, leading to an equal share of occurrences with the previous names. In German and even more in Dutch, *coronavirus* continued to be predominant, on the other hand. The more technical terms *n-cov* and *sars-cov-2* are only used marginally.

4.2 Most common associated words and their categories (RQ2)

The second research question concerns the most commonly associated words surrounding the main names for SARS-CoV-2 and the disease it causes (i.e., *virus*, *corona*, *coronavirus*, *covid*, *covid19*, *ncov*, *sars*, and *SARScov2*) in a cross-linguistic news media corpus. Table 3, shown in Section 4.1, illustrates the categories that resulted from our iterative coding procedure. We now proceed to a more quantitative perspective, assessing the distribution of the semantic categories.

Table 5 shows the categories with their numbers of associated words from the total set of 1,308. Both the absolute counts of tokens (number of occurrences *N*) and the relative counts (percentages) are provided. A chi-square analysis for the different categories and their numbers of items showed a significant relation [χ^2 (11) = 842.48, $p < 0.001$].

Read from bottom to the top, the most frequent category consists of terms that are part of the VISUAL & BIOMEDICAL FEATURES domain, demonstrating the importance of the medical description and scientific analysis of the virus and the disease. The next most frequent domains are MEASURES, highlighting specific consequences of the virus and disease as well as actions taken against it: DIFFUSION & SPREAD, highlighting inherent features of the pandemic; EVALUATION, representing an emotionally loaded expression; and INTENSITY, expressing degrees of strength, impact, or force. The categories that emerge least frequently from the data are COMPARISON, METALINGUISTIC DISCOURSE, (UN)CERTAINTY, CAUSE & EFFECT, LOCATION, RECENCY, and TEMPORAL SEQUENCE.

Table 5: Number and proportion of the 1,308 associated terms indexing a particular category, ordered by frequency of occurrence; color coded into categories that statistically group together

Semantic category	Associated words indexing the category	
	N **(out of 1,308)**	**%**
COMPARISON	53	4.05
METALINGUISTIC DISCOURSE	53	4.05
(UN)CERTAINTY	61	4.66
CAUSE & EFFECT	65	4.97
LOCATION	69	5.28
RECENCY	95	7.26
TEMPORAL SEQUENCE	95	7.26
INTENSITY	179	13.69
EVALUATION	204	15.60
DIFFUSION & SPREAD	218	16.67
MEASURES	251	19.19
VISUAL & BIOMEDICAL FEATURES	355	27.14

4.3 Cross-linguistic differences in the semantic domains (RQ 3.1, 3.3)

In order to address our third research question, we analyze the extent to which the distribution found in Table 5 shows variation over time across our target languages French, English, German, and Dutch.

A three-way log-linear analysis for the variables Month (Jan–Feb–March–Apr–May–June–July–Aug–Sept–Oct 2020), language (German, English, French, and Dutch), and semantic category ((UN)CERTAINTY, CAUSE & EFFECT, COMPARISON, MEASURES, DIFFUSION & SPREAD, EVALUATION, INTENSITY, LOCATION, METALINGUISTIC DISCOURSE, RECENCY, TEMPORAL SEQUENCE, and VISUAL & BIOMEDICAL FEATURES) shows a non-significant likelihood ratio (χ^2 (0) = 0, p = 1.00). The highest order interaction (Month × Language × Semantic category) was not significant (χ^2 (297) = 259.38, p = 0.944). The association between Language and Month was also not significant (χ^2 (27) = 13.88, p = 0.982).

By contrast, there was a significant association between semantic category and language [χ^2 (33) = 399.52, $p < 0.001$], illustrated in Figure 2. The heatmap shows the occurrence of the categories in relative frequency, calculating the number of associated words of 1,308 that index a specific category per language. Higher relative frequencies are reflected by darker colors. In the following, we only report on significant differences. The domains RECENCY and (UN)CERTAINTY appeared more frequently in German than in Dutch; for METALINGUISTIC DISCOURSE, this was the other way around. The frequency of EVALUATION and TEMPORAL SEQUENCE was higher for French than for English and German. The categories VISUAL & BIOMEDICAL FEATURES and LOCATION appeared more frequently in English than in Dutch and German, whereas the domain DIFFUSION & SPREAD appeared less frequently in French compared to that in Dutch and English. CAUSE & EFFECT was more frequent in Dutch than in English and French. In German, COMPARISON appeared more often than in the other languages. German and Dutch did not differ with respect to the frequency of INTENSITY, but these frequencies were higher than those for English and French. The category MEASURES did not show a significant difference across the languages.

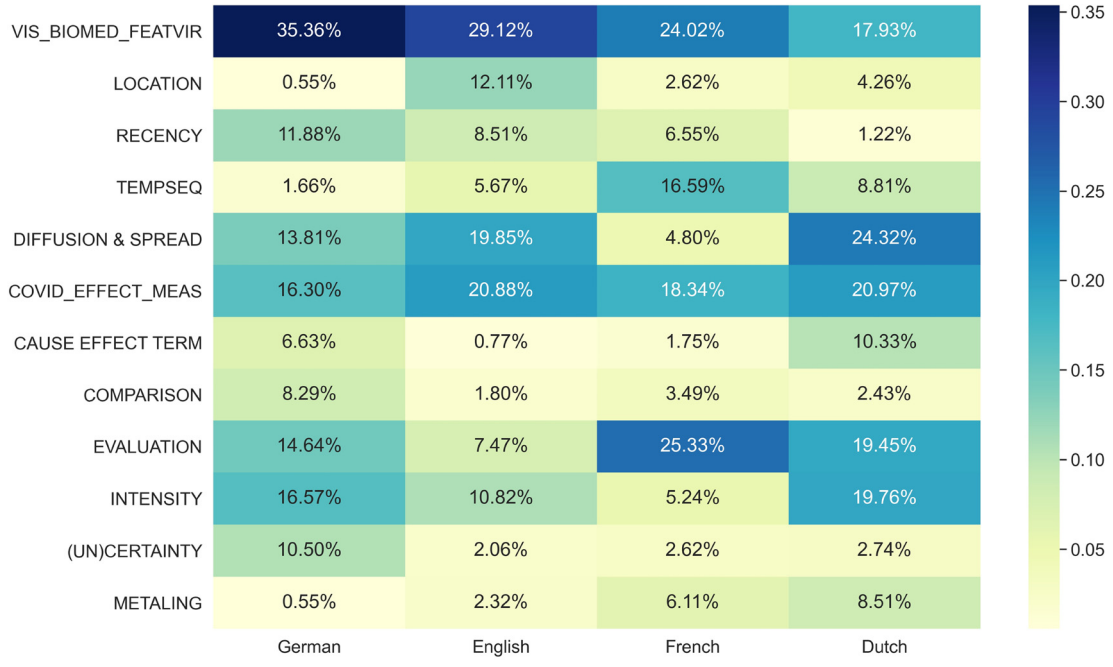

Figure 2: Heatmap of semantic categories (expressed in relative frequency) per language, across all months (contrastive perspective).

4.4 Changes in the semantic categories over the 10-month period (RQ 3.2, 3.3)

The earlier three-way log-linear analysis also showed a significant relation between semantic category and month (χ^2 (99) = 154.14, $p < 0.001$). The frequency of (UN)CERTAINTY, COMPARISON, and DIFFUSION & SPREAD decreased significantly after the first 2 months. CAUSE & EFFECT, TEMPORAL SEQUENCE, and MEASURES, on the other hand, appear more frequently after the first 2 months. For the following domains, there was no significant relationship between the associated words and month: METALINGUISTIC DISCOURSE, RECENCY, LOCATION, INTENSITY, EVALUATION, and VISUAL AND BIOMEDICAL FEATURES. The frequencies for the different domains and their distribution over time are shown in Figure 3.

An interpretation of these findings is provided in the following discussion.

5 Discussion

Our analysis of COVID-19 discourse in a 10-month cross-linguistic newspaper corpus has revealed four key insights. First, clear differences were found in the distribution of different competing names for the virus. After a period of strong lexical variation in each language, matters settled down differently in different languages. Where names relating to the virus itself (*corona, coronavirus*) are more popular in Dutch and German, names relating to the disease caused by the virus (*covid* and *covid19*) prevail in the English and French data.

Second, despite this variation in how the object of the health crisis is named, we mainly see cross-linguistic stability in the way it is being discussed. In all four languages under scrutiny, VISUAL & BIOMEDICAL FEATURES and EFFECTS & MEASURES are among the top three most frequently found semantic categories. Also, for each of the four languages, these more factual categories are complemented by semantic categories that provide an EVALUATION of the virus/disease or its INTENSITY. In this case, some differences are found between the languages, with French demonstrating a significantly higher tendency for EVALUATIVE terms than German and English. This can be seen in the frequent use of adjectives such as *satané* 'satanic', *maudit* 'cursed', *méchant* 'evil', and *mystérieux* 'mysterious'. These results closely resonate with the findings of Aslam et al. (2020) on negative polarity in COVID-19 newspaper headlines. In turn, it could be tempting to see evidence for a more epistemic or objective orientation in the English data, where the categories VISUAL & BIOMEDICAL FEATURES and LOCATION appeared

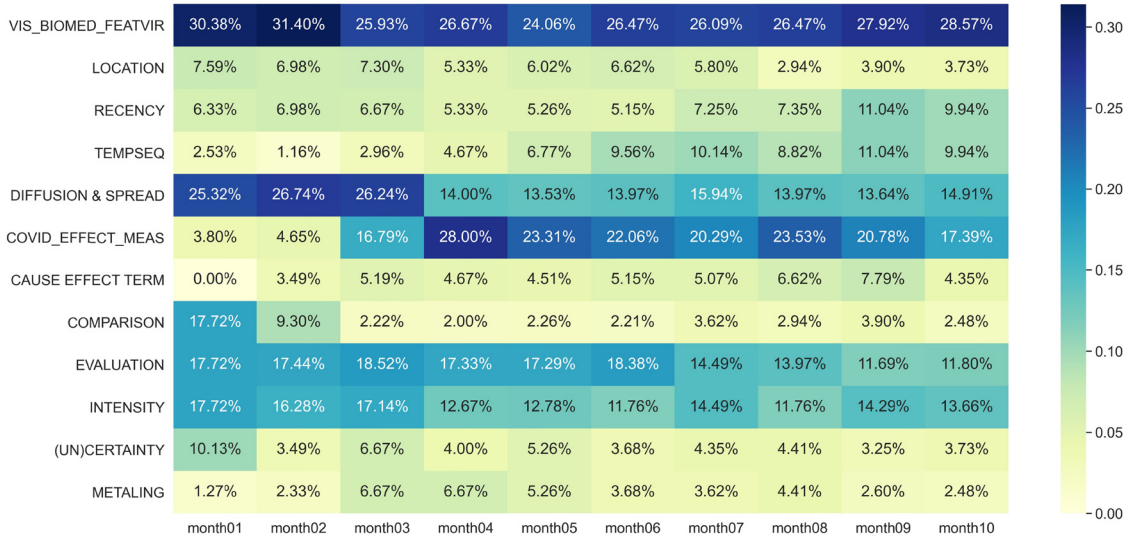

Figure 3: Heatmap of semantic category per month (relative frequency), across all languages (from month 01 = January to month 10 = October).

significantly more frequently than in German and in Dutch. This category, however, contains references to the origin of the first outbreak of the virus (*China, Wuhan*), which is by now often considered stigmatizing and hence surely not necessarily objective (see Budwhani and Sun 2020).

Third, no significant three-way interaction was attested between semantic category, language, and month. This provides further support for cross-linguistic stability in the description of the virus and the disease. In terms of chronological patterns, we first see stability across the four languages for the categories METALINGUISTIC DISCOURSE, RECENCY, LOCATION, INTENSITY, EVALUATION, and VISUAL AND BIO-MEDICAL FEATURES, which show no significant increase or decrease in our 10-month corpus. In contrast, a significant decrease was attested in the reporting of how the illness spreads subsequently to March 2020. It appears that, at that point, interest shifted to reporting on how to prevent and combat the illness through various MEASURES. After January 2020, there was a sharp decline in comparing COVID-19 to other illnesses (such as FLU and SARS viruses) concurrent with a drop in the amount of language devoted to (UN)CERTAINTY about the illness. It is noteworthy that across languages, the period from March 2020 marks a shift toward displaying more concrete MEASURES against the virus and how to start dealing with it. These findings support an interpretation that, regardless of region and language, it is of tantamount importance to first of all talk about what we are dealing with in a public health crisis, which is immediately followed by measures about how to prevent and combat it.

Before we arrive at final conclusions based on these patterns, let us note some methodological shortcomings in our treatment that attenuate our current analysis and could help shape follow-up studies. First, linguistic differences between the four languages of investigation hampered data retrieval and may have had an impact on the interpretations. Particularly, the constituent elements of compounds spelled as one word with the seed term were not included as potentially associated words. This is not an issue for French and English, where this type of compound is not a frequent (or often even plausible) outcome of derivational processes. For Dutch and German, where the spelling of compounds as one word as in *coronapatiënt* (Dutch) 'corona patient' or *Coronaregel* (German) 'corona rule' is a highly regular word formation process, this caused a loss of information. To accommodate for this shortcoming, a follow-up analysis identified the most frequent compounds for German and Dutch in our corpus that include our target seed terms. A first exploration reveals that the semantic categories identified in Table 3 succeed in capturing the meanings indexed by the associated words found in the compounds. Moreover, the lack of strong systematic differences between the German and Dutch subcorpora on the one hand and the French and English subcorpora on the other hand further mitigates the impact of this shortcoming in data collection.

Another aspect of concern is the diatopic distribution of web pages by country. This information would allow, for example, a further distinction between language varieties (e.g., Belgian Dutch vs Netherlandic Dutch, Austrian German vs German German, and French French vs Belgian French). However, this information is present in the metadata for only about 50% of the pages, so we have chosen to disregard this parameter in the present study. Future work could address this issue, as that will allow for assessing the interplay of cultural and linguistic factors more closely.

This final shortcoming brings us back to the implication of the central results of this article, viz., (1) the interlinguistic variation in how the object of the 2020 health crisis is named vs the fairly large interlinguistic stability, (2) in the semantic categories surrounding these names, and (3) in the diachronic shifts in the occurrence of these categories. Presumably, this stability should come as no surprise. We focus on languages spoken in and crossing the boundaries of four neighboring countries in Western Europe. Despite the occasionally outspoken differences in measures taken, the similarities in framing and responding to the crisis most likely outweigh the differences. Alluring though it may be to link up the few differences we have attested to cultural stereotypes, contrasting the more emotive French reactions with the more measure-oriented English and German patterns, this would be reading more into the results of collocation extraction than is warranted.

Instead, our methods and results have allowed us to uncover what is shared. Through our combination of corpus linguistic methods and semantic classification, we have arrived at a description of the underlining shared semantic categories that characterize COVID-19 discourse in the first 10 months of the pandemic, proposing a shared basis in the conceptualization and discursive treatment of the pandemic in French, English, German, and Dutch newspaper data.

6 Conclusion

As we have seen, more in-depth qualitative descriptions of the particularly associated terms in each of these languages form the first avenue for future research. Going beyond the aggregated data from the statistical analysis and taking into account the underlying corpus data will help further grasp similarities and differences between the languages under scrutiny.

Overall, the results of our study have indicated cross-linguistically shared tendencies in newspaper reporting on a sudden global crisis such as the COVID-19 pandemic while also highlighting some linguistic preferences and diachronic shifts. Through a combination of corpus linguistic methods and semantic classification, we have arrived at a description of the underlining semantic space that characterizes COVID-19 discourse in the first 10 months of the pandemic, proposing a shared basis in the conceptualization and discursive treatment of the pandemic in French, English, German, and Dutch newspaper data as illustrated in Figure 3. Future research could take a closer qualitative look at the associated words used across the languages and carve out language-specific forms of conceptualizing aspects of the pandemic.

Acknowledgments: The authors gratefully acknowledge the open access funding provided by the University of Klagenfurt.

Funding information: The authors state no funding involved.

Authors contributions: Data collection was carried out by the first author, EC, who carried out the technical corpus linguistic analysis. All authors contributed to the coding of the data. EZ, BH-S, UN, and FvM contributed to the analysis of the Dutch data. EC and EW-F contributed to the analysis of the French data. AO and EW-F contributed to the analysis of the German data. EP and GA contributed to the analysis of the English data. The quantitative data analysis was mainly in the hands of EC, BH-S, UN, FvM, and EZ. EC created all the figures in the article. BH-S, UN, FvM, and EZ created Tables 1–3 and 5. The qualitative data analysis was the final responsibility of AO and EW-F, who created Table 3. All authors contributed to the drafting of the manuscript.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analysed during the current study are available in the Zenodo, https://zenodo.org/record/7339615.

References

Abdul-Mageed, M., A. R. Elmadany, E. M. B. Nagoudi, D. Pabbi, K. Verma, and R. Lin. 2021. Mega-COV: A Billion-Scale Dataset of 100+ Languages for COVID-19. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 3402–20, Online. Association for Computational Linguistics. Doi: 10.18653/v1/2021.eacl-main.298.

Aslam, F., T. M. Awan, J. H. Syed, A. Kashif, and M. Parveen. 2020. "Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak." *Humanities & Social Sciences Communication* 7, 23.

Bai, G. H. 2020. "Fighting COVID-19 with Mongolian fiddle stories." *Multilingua* 39(5), 577–86.

Balnat, V. 2020. "*Unter Beobachtung*: Corona-Wortschatz im Deutschen und Französischen [Under observation: Corona-Vocabulary in German and French]." *Nouveaux Cahiers d'Allemand: Revue de linguistique et de didactique*. Association des Nouveaux Cahiers d'Allemand. (hal-02931171).

Belhaj, S. 2020. "La pandémie Covid-19 et l'émergence d'un nouveau technolecte [The Covid-19 pandemic and the emergence of new technological vocabulary]." *Revue Langues, Cultures et Sociétés* 6(1), 28–38.

Black, S. P. 2020. "Communicability, stigma, and xenophobia during the COVID-19 outbreak: 'common reactions?'" *Language, Culture and Society* 2(2), 242–51.

Blank, A. 1998. "Kognitive italienische Wortbildungslehre." *Italienische Studien* 19, 5–27.

Bowker, L. 2020. "French-language COVID-19 terminology: international or localized?" *The Journal of Internationalization and Localization* 7(1–2), 1–27.

Budhwani, H. and R. Sun. 2020. Creating COVID-19 stigma by referencing the novel coronavirus as the "Chinese virus" on Twitter: quantitative analysis of social media data. Journal of Medical Internet Research, 22, Article e19301, Doi: 10.2196/19301.

Chen, X. 2020. "Fighting COVID-19 in East Asia: the role of classical Chinese poetry." *Multilingua* 39(5), 565–76.

Chen, E., K. Lerman, and E. Ferrara. 2020a. *#COVID-19: The first public coronavirus twitter dataset*. Original date: 2020-03-15T17:32:03Z.

Chen, E., K. Lerman, and E. Ferrara. 2020b. "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set." *JMIR Public Health and Surveillance* 6(2), e19273.

Chun, C. W. 2020. "The return of the 'Yellow Peril' the fear of getting sick from the other." *Language, Culture and Society* 2(2), 252–9.

Cougnon, L.-A. and L. de Viron. 2020. "Covid-19 et communication de crise [Covid-19 and crisis communication]." *Focus linguistique sur les tweets francophones de Belgique*. https://osf.io/preprints/socarxiv/3qrcw/.

Craig, D. 2020. "Pandemic and its metaphors: Sontag revisited in the COVID-19 era." *European Journal of Cultural Studies* 23(6), 1025–32.

Davies, M. 2021. *The Coronavirus Corpus*. https://www.english-corpora.org/corona/.

Du, Y. 2020. "'I don't feel like talking about it'. Silencing the self under Coronavirus." *Language, Culture and Society* 2(2), 260–8.

Evert, S., P. Uhrig, S. Bartsch, and T. Proisl. 2017. "E-VIEW-affilation–a large-scale evaluation study of association measures for collocation identification." *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch*, p. 531–49.

Hu, Z., Z. Yang, Q. Li, and A. Zhang. 2020. "The COVID-19 infodemic: infodemiology study analyzing stigmatizing search terms." *Journal of Medical Internet Research* 22(11).

Katermina, V. and E. Yachenko. 2020. "Axiology of COVID-19 as a linguistic phenomenon in english mass media discourse." *Advances in Journalism and Communication* 8, 59–67.

Kilgariff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, et al. 2014. "The Sketch Engine: ten years on." *Lexicography* 1(1), 7–36.

Kurten, S. and K. Beullens. 2021. "#Coronavirus: Monitoring the belgian twitter discourse on the severe acute respiratory syndrome coronavirus 2 pandemic." *Cyberpsychology, Behavior, and Social Networking* 24(2), 117–22. Doi: 10.1089/cyber.2020.0341.

Ladilova, A. 2020. "Spanische Wortbildung im Kontext der Coronapandemie [Spanish wordformation in the context of the coronavirus pandemic]." In *Corona: Krise oder Wende? Wie Krisen Kulturen verunsichern und verändern* [*How Crises create insecurity and change cultures*], edited by M. O. Hertrampf, p. 44–55. Berlin: PhiN-Beiheft.

Leaman, R. and Z. Lu. 2020. "A comprehensive dictionary and term variation analysis for COVID-19 and SARS-CoV-2." *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Doi: 10.18653/v1/2020.nlpcovid19-2.32.

Lopez, C. E., M. Vasu, and C. Gallemore. 2020. "Understanding the perception of COVID-19 policies by mining a multilanguage twitter dataset." *ArXiv*. https://arxiv.org/abs/2003.10359.

Ma, H., L. Shen, H. Sun, Z. Xu, L. Hou, S. Wu, et al. 2021. "Database. COVID term: a bilingual terminology for COVID-19." *BMC Medical Informatics and Decision Making* 21, 231. Doi: 10.1186/s12911-021-01593-9.

Makhachashvili, R. and K. Bilyk. 2020. "L'analyse du corpus des textes 'covid-19' en utilisant le logiciel SketchEngine [Analyzing the Covid-19 text corpus using SketchEngine]." *Conference Paper. 30th of October 2020*. Strasbourg, France.

MIRCo. 2020. "Pandemic discourses and the prefiguration of the future." *Language, Culture and Society* 2(2), 227–41.

Nossem, E. 2020. "Linguistic rebordering: Constructing COVID-19 as an external threat." In *Borders in Perspective, Thematic Issue 4, Bordering in pandemic times, Insights into the COVID-19 Lockdown*, edited by C. Wille and R. Kanesu, p. 77–80. University of Luxembourg and University of Trier.

Parvin, G. A., R. Ashan, M. H. Rahman, and M. A. Abedin. 2020. "Novel coronavirus (COVID-19) pandemic: the role of printing media in asian countries." *Frontiers in Communication*. Doi: 10.3389/fcomm.2020.557593.

Pecina, P. 2010. "Lexical association measures and collocation extraction." *Language Resources and Evaluation* 44(1), 137–58.

Pietrini, D. 2020. "Non è distanza sociale! Parole nel turbine vasto [This is not social distancing! Words in the chaos]." *Treccani Magazine*, 29.04.2020.

Pietrini, D. 2021. *La lingua infetta. L'italiano della pandemia.* [*The infected language. The Italian language during the pandemic*]. Roma: Treccani.

Piller, I., J. Zhang, and J. Li. 2020. "Linguistic diversity in a time of crisis: language challenges of the COVID 19 pandemic." *Multilingua* 39(5), 503–15.

Rodríguez Abella, R. M. 2021. "Palabras para una pandemia. algunas notas sobre las creaciones neológicas utilizadas para comunicar la enfermedad por coronavirus SARS-CoV-2 [Words in the pandemic. Some notes on neologisms used to talk about coronavirus SARS-CoV-2]." In *Contribuciones a la Lingüística y a la Comunicación Social.* [*Contributions to linguistics and social communication*], edited by R. M. Rodríguez Abella, A. M. Alvarado, L. R. Miyarez, and L. Chierichetti, p. 78–82. Santiago de Cuba: Ediciones Centro de Lingüística Aplicada.

Roig-Marín, A. 2021. "English-based coroneologisms. a short survey of our Covid-19-related vocabulary." *English Today* 37(4), 193–5.

Rychlý, P. 2008. "A lexicographer-friendly association score." In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, 6–9.

Semino, E. 2021. "Not soldiers but fire-fighters – Metaphors and Covid-19." *Health Communication* 36(1), 50–8. Doi: 10.1080/10410236.2020.1844989.

Sgroi, S. C. 2020. *Dal Coronavirus al Covid-19. Storia di un lessico virale*. [From Coronavirus to Covid-19. The Story of a viral lexicon]. Alessandria: Edizioni dell'Orso.

Singh, L., S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, et al. 2020. "A first look at COVID-19 information and misinformation sharing on twitter." arXiv:2003.13907 [02.08.2020].

Solovejute, R. and D. Gatherer. 2020. Hunting for previous coronavirus pandemics using corpus linguistic analysis of 19th Century British Newspapers. *Preprints 2020, 2020120817*. Doi: 10.20944/preprints202012.0817.v1.

Spina, S. 2020. "Un confronto tra il discorso della stampa quotidiana e quello delle interazioni in twitter sul tema delle migrazioni [A comparison between daily news discourse and discourse on Twitter on the topic of migration]." In *Il discorso sulle migrazioni/Der Migrationsdiskurs* [*Migration discourse*], edited by D. Pietrini, p. 145–62. Frankfurt: Peter Lang.

Tan, K. H., P. Woods, H. Azman, I. H. Abdulah, R. Z. Hashim, H. A. Rahim, et al. 2020. "Covid-19 insights and linguistic methods. 3L: language, linguistics, literature." *The Southeast Asian Journal of English Language Studies* 26(2), 1–23.

Thiéry-Riboulot, V. 2020. "Une étude de sémantique historique du mot confinement [A study on the historical semantics of the word confinement]." *Mots. Les languages du politique* [*Words. The languages of politics*] 124, 127–44.

Trampus, M. and N. Blaz. 2012. "The internals of an aggregated web news feed." *Proceedings of 15th Multiconference on Information Society 2012* (*IS-2012*).

Wicke, P. and M. M. Bolognesi. 2020. "Framing COVID-19: how we conceptualize and discuss the pandemic on twitter." *PLoS One* 15(9), e0240010.

Zhang, J. and J. Li, Eds. 2020. "Linguistic diversity in a time of crisis: language challenges of the COVID-19 pandemic." *Special issue of Multilingua* 39(5).

Zhang, L.-T. and S. Zhao. 2020. "Diaspora micro-influencers and COVID-19 communication on social media: the case of Chinese-speaking youtube vloggers." *Multilingua* 39(5), 553–63.

Zhu, H. 2020. "Countering COVID-19-related anti-Chinese racism with translanguaged swearing on social media." *Multilingua* 39(5), 607–16.