Humanitarian reports on ReliefWeb as a domain-specific corpus

Loryn Isaacs

Department of Translation and Interpreting, EDHCSJ,
University of Granada, Buensuceso, 11, 18002 Granada (Spain)
E-mail: lisaacs@ugr.es
https://orcid.org/0000-0003-0267-4853

Abstract

This paper presents an assessment of the content available on ReliefWeb's API for its suitability as a domain-specific corpus. ReliefWeb's position as a primary information resource for humanitarian response, boasting a database of nearly a million reports, lends it considerable value for the corpus-based study of humanitarian discourse. However, the service's content is under-explored in this regard. To this end, a Python package is introduced to manage the creation of ReliefWeb corpora. The composition of ReliefWeb's HTML reports in English is examined and compared with a corpus from the Humanitarian Encyclopedia. The comparison includes a keyness analysis of the Encyclopedia's 129 concepts and an assessment of diachronic trends for six concepts (HUMANITARIAN REFORM, SUSTAINABILITY, RESILIENCE, GENDER-BASED VIOLENCE, SETTLEMENT, and SOVEREIGNTY), as well as an analysis of hypernymic and definitional knowledge-rich contexts. Results indicated that ReliefWeb reports, mostly brief news and press release items, have much lower relative frequencies for humanitarian concepts than the reference corpus. Still, the data overlapped considerably and the breadth of the HTML content contributed important thematic diversity for some concepts. The paper concludes with a discussion of how the management of ReliefWeb corpora could be improved in future iterations.

Keywords: humanitarian domain; corpus creation; ReliefWeb; information extraction

1. Introduction

ReliefWeb[1] is a service managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) that aggregates publicly available documents related to current humanitarian issues. In 2023, this 27-year-old database is likely to reach one million reports spanning half a century and representing thousands of diverse actors worldwide. During the first half of 2022 it had 10.8 million users and experienced increases across a variety of usage metrics (OCHA, 2022). In fulfilling its founding objective to "act as the principal information system for prevention, preparedness, and rapid response for the humanitarian community" (Ruso, 1996, p. 18), the service also represents a significant resource in the study of humanitarian communication. The database has been utilized in various fashions, such as in

[1] https://reliefweb.int/

ReliefWeb Labs projects,[2] to discursively track famine (Rubin, 2014), and to extract knowledge via semantic embedding (Shamoug, Cranefield & Dick, 2023). A common goal is to improve humanitarian response by leveraging linguistic data, which can be hampered by the difficulty of synthesizing and transmitting domain knowledge.

This paper approaches the database from a corpus-based linguistics perspective with several aims. One is to provide a thorough analysis of ReliefWeb's composition, which can be treated superficially despite its relevance in guiding data interpretation. Another is to convert the bulk of ReliefWeb reports into a format readable by language corpus management software and to offer a means to periodically update such corpora. This is with the hope of establishing an accessible and durable means to facilitate research in regards to humanitarian knowledge extraction and representation. Finally, this article assesses the suitability of ReliefWeb's HTML reports for analyzing domain-specific concepts. While the database's value is apparent, understanding its limits as a representation of humanitarian discourse is a necessary practice.

The assessment of ReliefWeb data provided here is part of ongoing efforts to expand and refine the corpus-based methods used to generate concept entries for the Humanitarian Encyclopedia platform.[3] The Encyclopedia, a project by the Geneva Centre of Humanitarian Studies, offers analyses on 129 humanitarian concepts with a combination of corpus-based linguistic reports and input from domain experts. It documents aspects of humanitarian discourse with a focus on concept variation and multidimensionality (León-Araúz, 2017), and also promotes community discussion of the domain's lexicon. To develop entries, data are retrieved with semantic and multiword-term querying techniques, and a battery of visualizations are supplied to ground discussion quantitatively (Chambó & León-Araúz, 2021; León-Araúz & San Martín, 2018). Concept analyses for the Encyclopedia have so far been conducted on an internal corpus of 4,824 public humanitarian documents from the last two decades. However, a significant portion of this corpus is likely included in ReliefWeb, which boasts both a mature data management system and continuously updated content. While leveraging the service's content is a logical progression, this requires converting ReliefWeb's reports into a tokenized, lemmatized language corpus and studying its shape and limitations.

Section 2 describes an API-based data retrieval method and the conversion of data into a Sketch Engine-compatible format (Kilgarriff et al., 2014). The section summarizes the ReliefWeb corpus's composition and describes the methodology used to compare it with the Humanitarian Encyclopedia's corpus. Section 3 reports results for a keyness analysis of 129 humanitarian concepts, as well as results regarding diachronic trends for six concepts and the density of their hypernymic and

[2] https://labs.reliefweb.int/

[3] https://humanitarianencyclopedia.org/

definitional contexts. Section 4 discusses how the results pertain to the use of ReliefWeb as a domain-specific corpus.

2. The ReliefWeb English corpus

2.1 Data collection, structure and limits

ReliefWeb reports are one of several content types available via API request to the service. The category contains the bulk of the site's primary content: an empty query returned over 988,000 results as of April 2023, including documents, maps, and other digital formats aggregated from internet sources. The corpus described in the following sections focuses on reports in English with an original publication date from 2000 to 2022, excluding several text-poor formats, namely maps, interactives, and infographics, as well as the heterogeneous "other" category. This returned 662,473 API responses (67% of reports at the time).

The above figures require some contextualization. Importantly, many of the report metadata fields (a total of 98 were detected) allow multiple values. Language is one of these, meaning that a request for English content includes any report with at least a tag for English. Reports of this type may have multiple texts with different languages, as in the uncommon case where several PDF translations are available; just under 2% of the collected reports contained other language tags in addition to English. Tagging errors can also introduce some non-English texts in the data recovered (e.g., as of publication, report no. 21366).[4] While these sources of noise appear minimal, employing language detection algorithms would likely be necessary to establish more exact figures.

Defining what "report" means in ReliefWeb's database is also a prerequisite. Each text available through the API[5] is given a unique identifier. Each identifier refers to at least one form of content, but a report may be a collection of related materials. The content visible on ReliefWeb's website is all of or a portion of what is understood as the report's primary document. For shorter texts, like press releases, most or all body text (e.g., excluding footers) may be displayed as HTML content. For longer texts, however, only a portion is displayed, such as a document's introduction, executive summary, or first page.

An example report is given below for a publication from Humanitarian Practice Network (web page content is in Figure 1 and the source PDF in Figure 2).[6] In this case, the HTML text on ReliefWeb contains 366 words: this is a portion of the PDF's

[4] https://reliefweb.int/node/21366

[5] https://api.reliefweb.int/v1/reports

[6] https://reliefweb.int/node/23456

first page, with one of the middle paragraphs removed (starting with "The articles in the special feature"). In other words, a report may consist of a portion of altered text that meets ReliefWeb's editorial constraints for size and content. In this example the full PDF is 52 pages, while the searchable text is the report's HTML body.

Figure 1: Report no. 23456 HTML content

Figure 2: Report no. 23456 original PDF

Report, then, is used here to refer to each unique item in the database and, more specifically, the HTML content for these items that users view when browsing (the *body-html* API field). Since a corpus built from this data excludes full-text PDF content, which exists for nearly a third of the downloaded reports, full-text analysis is

not possible. Conversely, since two thirds of reports have no PDF data, HTML-only content may be more complete, albeit for genres with shorter average lengths.

In addition to the aforementioned architectural limitations, authors have noted concerns familiar to corpus linguists. As one of the service's founders states, "Information is not neutral. The user must judge the reliability of content and the biases in reporting" (Ruso, 1996, p. 120). While the methods and principles guiding the service have developed over the course of decades, some factors regarding neutrality or bias may still be relevant. The first is perhaps the primacy of English language texts, a challenge recognized in past recommendations (Naidoo, 2007, p. 57). The future also holds new issues for the online data aggregation service, including disinformation (Wackernagel & Footner, 2021).

Aspects of data collection aside, authors have pointed to several considerations for interpreting ReliefWeb's linguistic data. One is that content published during and immediately after emergencies may suffer in quality, originality, and substance (von Schreeb et al., 2013). In other words, while quality issues with crisis reporting may be corrected in successive documents, initial errors may remain in the corpus as artifacts that could later skew results. Other fundamental concerns for approaching the domain's discourse include humanitarian concepts lacking standardization; poor contextualization of term frequencies; politically and institutionally motivated uses and omissions; changes in the distribution and representation of organizations; under- and over-reporting of geographic areas due to accessibility; and data reported on national levels obscuring local trends (Rubin, 2014; von Schreeb et al., 2013).

2.2 Corpus compilation

After JSON API response data was flattened and stored in an SQLite table, *html-body* text was processed with a Stanza NLP pipeline utilizing the default Universal Dependencies English Web Treebank (EWT) model (Qi et al., 2020; Silveira et al., 2014). Output was reshaped into a vertical format, with an XML string containing metadata inserted into each text. Given that 98 metadata tags were detected in the data set, only those judged most valuable for corpus queries were included, 22 in total. Discarded tags include country coordinates, URLs to associated images, and redundant categories (*country.iso3* being preferred over *country.id*).

Since many ReliefWeb metadata fields allow for multiple values, fields with lists of values were concatenated into strings with a pipe separator. For example, a report with multiple values in *source.name* appears as "World Health Organization|Government of Nigeria", and the corresponding values in related tags maintain the same list order, with *source.type.name* being "International Organization|Government"). The original data structure is maintained in this way,

although it can cause confusion if tags are viewed in isolation, as some include over thirty values, e.g., report no. 630723.[7]

A tagset file was produced by detecting unique XPOS values (52) generated from the Stanza pipeline. This tagset was compared with Sketch Engine's default tagger,[8] a modified version of TreeTagger (Marcus et al., 1993), to identify dissimilarities that end users should take into account. Most of the EWT and modified TreeTagger tags were functionally equivalent, with the exception of verb tags, given that Sketch Engine has a separate set of tags exclusively for the verb *be*. When designing queries, Stanza's more atomic tokenization for hyphenation should be taken into account: for example, "gender-based" gets split into three tokens rather than one.

Compressed archives of vertical text were then exported and fed to Sketch Engine's corpus compilation tool. Specifically, this took the form of a local NoSketch Engine server run as a Docker container (Kilgarriff et al., 2014; National Laboratory for Digital Heritage, Eötvös Loránd University Department of Digital Humanities, 2023; Rychlý, 2007). The Python functions to replicate data retrieval and corpus creation have been made available in a GitHub repository (Corpusama, v0.1.1).[9] The workflow was configured to update the corpus by last date of modification, meaning that prior versions of content could be overwritten without being tracked if a report is updated post-publication.

2.3 Corpus composition

In total, 662,473 API responses produced 431,170,905 tokens, 366,049,459 words, 16,809,660 sentences, and 657,098 documents, with the last figure being slightly lower than the total number of API responses due to some reports lacking any text content (i.e., only containing other HTML elements). The average document length amounted to 557 words. Table 1 summarizes corpus attributes, ordered by the number of unique items and combination of items ("structure frequency" in Sketch Engine). For example, among the 47,152 values for *country.shortname* are "Israel", "oPt", "Israel|oPt", "oPt|Israel", and other combinations with over a dozen countries. For this attribute a total of 248 countries are represented, meaning that most of its values are combinations of several countries. This multivalue format preserves ReliefWeb's data structure, but it also inflates tallies for some attributes: "Israel|oPt" and "oPt|Israel" may have no meaningful difference but are nonetheless counted separately.

[7] https://reliefweb.int/node/630723

[8] https://www.sketchengine.eu/english-treetagger-pipeline-2

[9] https://github.com/engisalor/corpusama

Attribute	Items [a]	Unique [b]	NA% [c]	Example [d]
id	657,098	657,098	0	100001
url	657,098	657,098	0	https://reliefweb.int/node/100001
title	652,358	652,358	0	Food Security Outlook Update May2011
origin	381,773	381,773	38	https://www.ifrc.org/appeals
country.iso3	47,152	248	0	afg
country.shortname	47,152	248	0	Afghanistan
theme.name	14,492	21	12	Protection and Human Rights
disaster.name	13,041	2,621	62	Haiti: Earthquakes - Jan 2010
disaster.glide	12,888	2,508	61	OT-2011-000205-NER
source.name	11,155	2,708	< 1	The New Humanitarian
source.shortname	11,133	2,684	< 1	TNH
source.homepage	10,884	2,485	< 1	http://www.unhcr.org/
date.original	8,395	8,395	0	2017-06-30T00:00:00+00:00
source.spanish_name	3,848	197	48	Gobierno de Filipinas
disaster_type.name	2,436	22	45	Tropical Cyclone
source.type.name	1,122	8	< 1	International Organization
primary_country.iso3	235	235	0	wld
primary_country.shortname	235	235	0	World
date.original.year [e]	23	23	0	2017
language.name	17	5	0	English
ocha_product.name	14	14	97	Flash Update
format.name	9	9	< 1	News and Press Release

[a] Includes individual items (*Ethiopia*) and lists (*Ethiopia|Kenya* and *Kenya|Ethiopia* being distinct).
[b] Includes individual items only (*Ethiopia*, *Kenya*, *Somalia*).
[c] Percentage of *NA* values in the total frequency: 0 = no missing values; < 1 is a non-zero result.
[d] Examples from various reports.
[e] Extracted from date.original during compilation.

Table 1: ReliefWeb corpus attributes

Table 2 offers further details on corpus composition, showing the top ten values for several attributes. Frequencies and relative frequencies refer to the total number of instances of an item: "World" occurs 65,395 times in *county.shortname*, whether alone (29,100) or as part of lists, including "Greece|World" (608), "Libya|World" (545), etc. Altogether, "World" appears in 16,209 different lists, almost 98% of which have five or fewer occurrences. This long tail is characteristic of *country.shortname* and similar ReliefWeb attributes.

Attribute	Value	*frq*	*relfreq*
country.shortname	World	65,395	151.67
	Sudan	46,418	107.66
	Afghanistan	42,672	98.97
	Syria	37,781	87.62
	DR Congo	34,948	81.05
	Somalia	34,820	80.76
	Iraq	34,497	80.01
	oPt	33,025	76.59
	Pakistan	30,423	70.56
	Ethiopia	27,862	64.62
date.original.year	2020	37,328	86.57
	2015	35,816	83.07

Attribute	Value	*frq*	*relfreq*
	2009	34,942	81.04
	2022	34,690	80.46
	2017	34,532	80.09
	2014	33,784	78.35
	2018	33,446	77.57
	2019	32,839	76.16
	2016	32,210	74.7
	2021	31,845	73.86
disaster_type.name	NA	403,656	936.19
	Flood	80,114	185.81
	Epidemic	75,544	175.21
	Drought	48,386	112.22
	Earthquake	40,108	93.02
	Tropical Cyclone	39,872	92.47
	Land Slide	30,945	71.77
	Flash Flood	28,507	66.12
	Other	19,367	44.92
	Drought|Other	14,317	33.2
format.name	News and Press Release	456,371	1058.45
	Situation Report	126,377	293.1
	Analysis	37,992	88.11
	Assessment	11,993	27.81
	Appeal	7,055	16.36
	Manual and Guideline	6,823	15.82
	UN Document	5,527	12.82
	Evaluation and Lessons Learned	4,792	11.11
	NA	168	0.39
source.name	The New Humanitarian	30,726	71.26
	UN High Commissioner for Refugees	27,245	63.19
	UN Office for the Coordination of Humanitarian Affairs	25,684	59.57
	World Health Organization	24,030	55.73
	World Food Programme	23,567	54.66
	Reuters - Thomson Reuters Foundation	22,614	52.45
	UN Children's Fund	22,317	51.76
	International Federation of Red Cross And Red Crescent Societies	18,288	42.41
	International Organization for Migration	14,273	33.1
	UN News Service	11,955	27.73
source.type.name	International Organization	277,390	643.34
	Media	115,298	267.41
	Non-governmental Organization	114,691	266
	Government	102,135	236.88
	Red Cross/Red Crescent Movement	34,065	79.01
	Academic and Research Institution	20,744	48.11
	International Organization|International Organization	11,902	27.6
	Other	9,200	21.34
	Government|International Organization	4,025	9.34
	International Organization|Government	2,084	4.83
theme.name	Protection and Human Rights	183,709	426.07
	NA	183,678	426

Attribute	Value	*frq*	*relfreq*
	Health	172,045	399.02
	Food and Nutrition	142,468	330.42
	Water Sanitation Hygiene	97,461	226.04
	Shelter and Non-Food Items	89,913	208.53
	Agriculture	70,714	164
	Education	60,230	139.69
	Contributions	55,310	128.28
	Coordination	53,564	124.23

Table 2: ReliefWeb corpus text type analysis

Frequencies for several key attributes yield some of the corpus's general characteristics:

1. The top countries have relatively comparable frequencies, with the highest counts ranging between 46,418 (3.2%) for Sudan and 16,536 (1.1%) for Haiti.

2. Though an increase in annual document counts is expected, no single year between 2000 and 2022 is a particular outlier.

3. While many disasters are not categorized by type, *flood* and *epidemic* are the most common.

4. Over two thirds of the corpus consist of news and press releases, with another fifth being situation reports.

5. While there are 2,708 contributing organizations, almost 30% of reports originate from the top ten sources, led by The New Humanitarian (4.1%).

6. Most sources fall under international organizations (39%), media (16%), NGOs (16%), and governments (15%).

7. Almost 12% of documents lack a theme. A considerable portion with themes refer to protection and human rights (12%), health (11%), food and nutrition (9%) and water, sanitation and hygiene (6.2%).

In brief, in the first 22 years since the turn of the century, ReliefWeb posted an average of close to 17 million words in English annually. This consists mainly of short news and press release items tagged for natural disasters, with potentially edited content that provides at least a document's summary. The most commonly tagged countries are from the African and Eastern Mediterranean World Health Organization regions[10] and are often grouped into wider affected areas. Among thousands of authors, led by international organizations, a small subset provides a

[10] https://www.who.int/countries

substantial amount of content, particularly those affiliated with the United Nations.

2.4 Assessment methodology

After compiling the corpus of ReliefWeb reports, an initial assessment was conducted to compare its characteristics against the corpus developed by the Humanitarian Encyclopedia (the HE corpus). The primary concern was how the low average word count of HTML texts on ReliefWeb (557) could affect the frequencies of the 129 humanitarian concepts studied by the Humanitarian Encyclopedia, given that the HE corpus is made up of PDFs averaging 14,760 words (over 26 times longer). The main objective of the following analysis, then, was to determine whether ReliefWeb's curated HTML content would be suitable for the Encyclopedia's concept analyses in lieu of compiling a complete corpus of the service's PDFs.

In comparison with ReliefWeb's multitudinous reports, the HE corpus is a much smaller set of publicly available humanitarian documents (4,824 texts amounting to 71,201,157 words) published between 2005 and early 2019. As both corpora consist of documents published online by humanitarian organizations, much of their content is expected to overlap, although to what extent is unknown without aligning their metadata. Both corpora have tags for document format, organization type, and geographic region, but these are not necessarily comparable. For example, geographic metadata in ReliefWeb refer to individual affected countries, whereas in the HE corpus they refer to the continent a document originated from. Given limitations of this nature, the current analysis considered only year of publication as a viable text type for comparison.

The analysis was conducted in three steps. First, the frequencies of the Humanitarian Encyclopedia's 129 concepts were collected from each corpus via a Python-based NoSketch Engine API controller (Isaacs, 2022). Corpus Query Language (CQL) rules (Jakubíček et al., 2010) were designed for these queries, most being uncomplicated (`[lemma_lc="knowledge"]` for KNOWLEDGE), while others took into account hyphenation, part-of-speech, or common abbreviations. For instance, the rule for INTERGOVERNMENTAL ORGANIZATION was as follows:

```
( ( [lemma_lc="inter-governmental|intergovernmental"] |
    [lemma_lc="inter"] [lemma_lc="governmental"] |
    [lemma_lc="inter"] [lc="-"] [lemma_lc="governmental"]
  ) [lemma_lc="organisation|organization"]
) | [lc="IGOs?"]
```

To compare the density of concepts across the corpora, normalized frequencies were used to compute an effect-size keyness score (Gabrielatos, 2018; Kilgarriff, 2012), which indicates whether a concept is more common in the focus corpus (ReliefWeb, or simply RW) or reference corpus (Humanitarian Encyclopedia, HE). The

distribution of keyness was analyzed with the assumption that many scores gravitating toward K=1 indicates a shared focus for a concept. In contrast, consistently low or high keyness across the 129 concepts could indicate important dissimilarities for the Humanitarian Encyclopedia's concept analyses.

Second, concepts were assessed by visualizing their frequency over time with the *DATE* text type for HE and *date.original.year* for RW. This utilized the *reltt* measurement in Sketch Engine (relative text type frequency), which is a per million tokens calculation that normalizes frequencies for text type values (in this case, each individual year). However, despite the fact that the corpora both focus on the same time period, they are not fully comparable. RW includes more years (23 versus 14) and HE has some tagging irregularities: documents missing a year (*DATE*=0), multiyear tags (*DATE*=2005-2006), and incomplete data for 2019. With these caveats in mind, comparing the data by year was intended to add perspective to the keyness analysis and identify whether the corpora displayed similar trends for concept frequencies over the first two decades of the century.

Third, six concepts with a range of keyness and which displayed varying diachronic trends (downward, upward, stable) were selected for an analysis of their hypernymic and definitional contexts. Random samples of 1,000 concordances were inspected manually in each corpus to compare the density and diversity of these knowledge-rich contexts, or KRCs (Condamines, 2022; Marshman, 2022; Meyer, 2001). This followed the Humanitarian Encyclopedia's concept analysis procedure, which is informed by Frame-based Terminology (Faber, 2022) and utilizes KRC-based knowledge extraction techniques (León-Araúz & San Martín, 2018; San Martín et al., 2020).

As part of this analysis, lists of monolexical hypernyms were collected and compared to judge the RW corpus's potential for knowledge extraction. KRCs with polylexical hypernyms were simplified to facilitate this comparison. For example, in the phrase "Resilience is also a contested term in the literature" in the RW corpus, TERM was extracted as the hypernym, with the (quite valuable) adjective "contested" being left for future discussion. Definitional contexts were identified with a flexible approach to maximize the number of available candidates. This allowed for formal and informal definitions containing genus and differentiae, verbal patterns, or paralinguistic patterns (Meyer, 2001; Sierra et al., 2008).

The number and qualities of the hypernyms and definitional contexts were compared; results were then discussed regarding the content for humanitarian concepts in RW and HE. Results followed the Humanitarian Encyclopedia's concept entry structure, which manages polysemous terms under a single entry. In other words, although `[lemma_lc="settlement"]` retrieved hypernyms referring to both SETTLEMENT=COMMUNITY (a population inhabiting a geographic area) and SETTLEMENT=BARGAIN (a mutual agreement), these were grouped together.

3. Results

3.1 Keyness for humanitarian concepts in ReliefWeb

Table 3 and Table 4 display the keyness and normalized frequency in RW by quartile for each of the Humanitarian Encyclopedia's 129 concepts. Keyness ranged from 0.006, for RIGHT-BASED APPROACH, to 3.850, for SOVEREIGNTY, with Q1=0.437, Q2=0.693, and Q3=1.060. The least frequent concept in absolute terms for RW was also RIGHT-BASED APPROACH, with 39 occurrences, compared to a maximum of 876,392 occurrences for CHILD (K=0.940). The large majority of concepts were less common in RW (K<1), with 93 concepts or 72%. 38 concepts were at least half as common in RW as HE (K<0.5); 10 were at least twice as common in RW (K>2).

Q1			Q2		
concept	***K***	***fpm***	**concept**	***K***	***fpm***
right-based approach	0.006	0.09	humanitarianism	0.438	1.83
logistic	0.027	0.96	inclusion	0.440	40.93
equity	0.112	12.29	participation	0.446	118.92
remote-sensing	0.122	0.16	private sector	0.451	70.82
humanitarian reform	0.143	1.18	diplomacy	0.480	13.47
advocacy	0.177	56.96	program	0.505	1,208.59
urbanisation	0.229	7.93	community-based approach	0.517	1.96
efficiency	0.245	23.93	cash	0.530	194.94
innovation	0.265	31.54	integrated approach	0.535	6.55
humanitarian action	0.273	32.69	aid dependence	0.536	1.04
sustainability	0.274	29.26	education	0.558	599.00
knowledge	0.279	86.64	humanitarian space	0.558	3.83
empowerment	0.281	34.68	context	0.566	156.18
effectiveness	0.302	39.19	mitigation	0.576	40.04
competition	0.304	17.90	adaptation	0.583	44.74
governance	0.309	95.06	datum	0.583	253.48
policy	0.337	357.08	empathy	0.589	1.99
grand bargain	0.346	3.67	dignity	0.592	52.25
management	0.346	403.67	leadership	0.596	118.09
partnership	0.347	192.63	care	0.598	483.37
capacity-building	0.369	58.73	localisation	0.598	3.88
do no harm	0.380	1.54	resilience	0.603	126.25
acceptance	0.381	12.74	corruption	0.617	46.21
technology	0.382	86.48	ethics	0.630	4.30
quality	0.390	155.98	intervention	0.638	201.79
disaster risk reduction	0.392	52.64	climate change	0.662	180.73
leave no one behind	0.402	4.73	service	0.664	920.50
development	0.402	1,122.52	risk	0.672	602.80
International governmental organisation	0.414	3.04	neutrality	0.672	11.10
poverty	0.418	216.56	prevention	0.676	178.47
accountability	0.429	89.47	implementation	0.685	300.41
culture	0.434	50.45	nutrition	0.693	220.27
local organisation	0.437	10.89			

Note. Smoothing = 0. Reference corpus = Humanitarian Encyclopedia. *fpm* = frequency per million tokens (focus corpus).

Table 3: Keyness of humanitarian concepts in ReliefWeb (lower half)

Q3			Q4		
concept	**K**	**fpm**	**concept**	**K**	**fpm**
vulnerability	0.694	107.41	response	1.091	919.16
faith	0.697	27.51	emergency	1.093	957.59
funding	0.706	317.57	coordination	1.097	341.15
humanitarian-development nexus	0.708	0.80	gender-based violence	1.135	80.05
impact	0.735	420.77	recovery	1.165	226.77
communication	0.751	195.08	early warning	1.181	59.35
evidence	0.765	107.06	food security	1.241	231.82
impartiality	0.765	12.82	testimony	1.264	17.50
community	0.795	1,662.05	power	1.293	256.40
justice	0.796	215.03	independence	1.319	78.45
aid	0.802	819.64	security	1.335	1,447.80
community engagement	0.814	11.32	need	1.356	1,183.39
monitoring	0.814	188.13	authority	1.361	629.82
solidarity	0.832	61.41	conflict	1.370	1,043.01
rehabilitation	0.833	178.78	armed actors	1.381	4.86
humanity	0.835	69.21	forced displacement	1.460	15.75
politics	0.859	25.08	needs assessment	1.479	34.94
humanitarian imperative	0.871	1.55	mandate	1.537	167.50
non-governmental organisation	0.871	419.32	epidemic	1.589	86.30
health	0.874	1,810.05	settlement	1.837	207.07
protection	0.884	580.74	crime	1.846	227.17
psychosocial support	0.889	43.00	negotiation	1.921	116.30
contingency planning	0.912	7.97	peace	2.057	844.76
child	0.940	2,032.59	affected population	2.293	65.60
access	0.944	704.60	shelter	2.296	426.33
sanitation	0.945	278.86	famine	2.352	67.74
humanitarian actor	0.947	35.61	terrorism	2.446	79.43
livelihood	0.954	261.32	humanitarian worker	2.708	29.82
transition	1.029	102.09	civilian	3.081	667.50
law	1.044	515.61	responsibility-to-protect	3.109	14.30
crisis	1.051	560.05	evacuation	3.326	89.44
ethnicity	1.060	13.71	sovereignty	3.850	31.76

Note. Smoothing = 0. Reference corpus = Humanitarian Encyclopedia. *fpm* = frequency per million tokens (focus corpus).

Table 4: Keyness of humanitarian concepts in ReliefWeb (upper half)

3.2 Diachronic change in humanitarian concept frequencies

As nearly three quarters of the concepts had K<1, most visualizations were similar to the one below for SUSTAINABILITY. These displayed generally flat or upward trending distributions across time for both corpora and higher relative frequencies for HE. An exception was HUMANITARIAN REFORM, one of the few concepts with a marked decline in use. Many of these graphs were punctuated by outlying values for problematic HE date tags (0 for missing values, 2019 being incomplete, and multiyear values like 2004-2005). Whereas typical years in HE have between 109 and 622 documents each, multiyear values each only appear once and hence were excluded below.

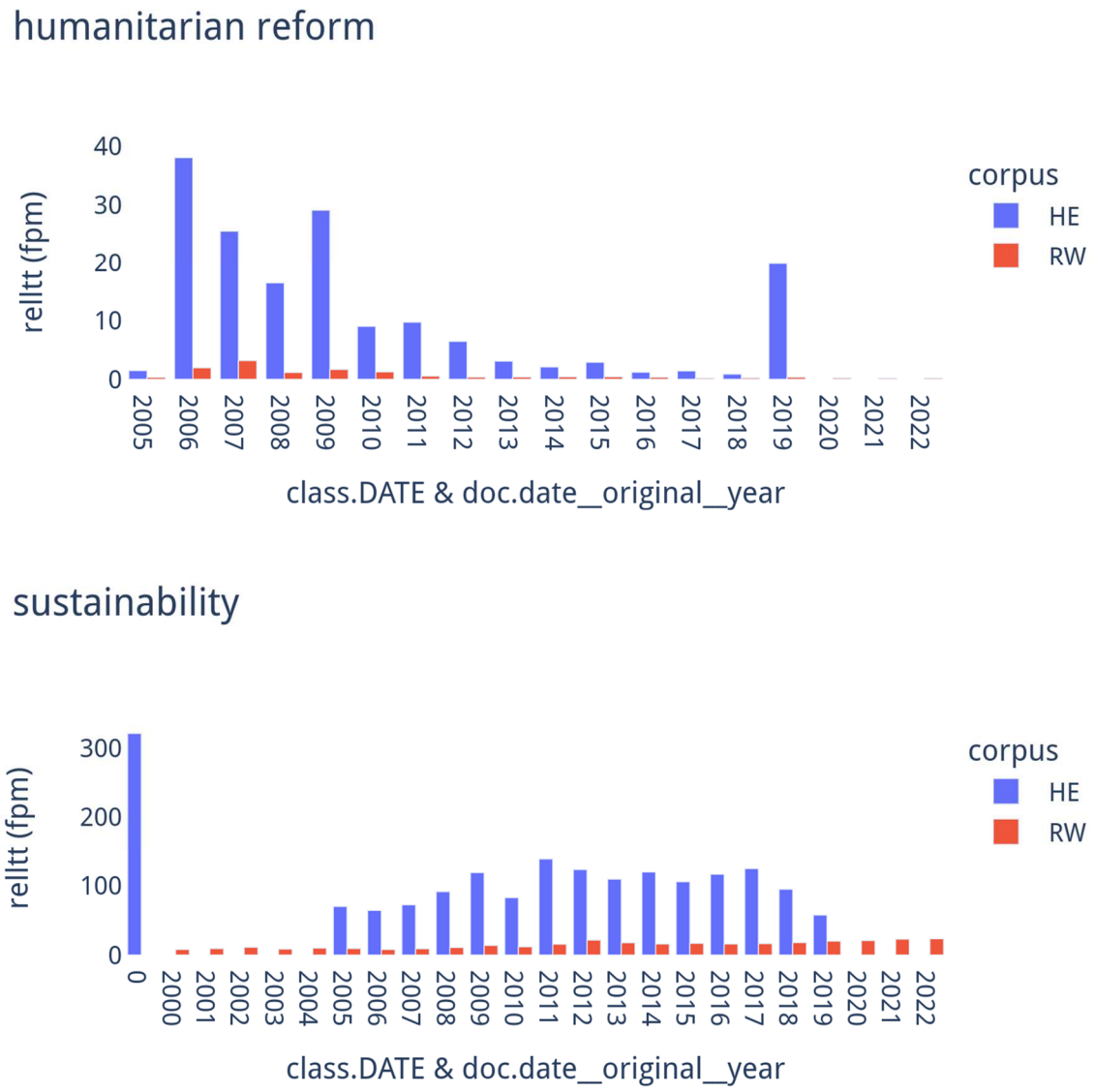

Figure 3: Concepts with K<Q1

The additional years covered in RW that HE lacks inflated the corpus-wide keyness for concepts. When keyness was computed for each shared, complete year and then averaged, scores dropped by half: the adjusted quartiles were Q1=0.232, Q2=0.358, and Q3=0.552 (compared to Q1=0.437, Q2=0.693, and Q3=1.06), with a maximum keyness of 2.199 for RESPONSIBILITY-TO-PROTECT. With these data, only ten concepts had K>1 (PEACE, CRIME, SHELTER, HUMANITARIAN WORKER, TERRORISM, AFFECTED POPULATION, CIVILIAN, SOVEREIGNTY, EVACUATION, RESPONSIBILITY-TO-PROTECT), with the final two above K>2. In contrast, 87 concepts had K<0.5, of which 38 had K<0.25. As seen in Figure 4, the annual relative frequencies of common concepts in RW often matched those of HE or were lower than suggested by corpus-wide keyness.

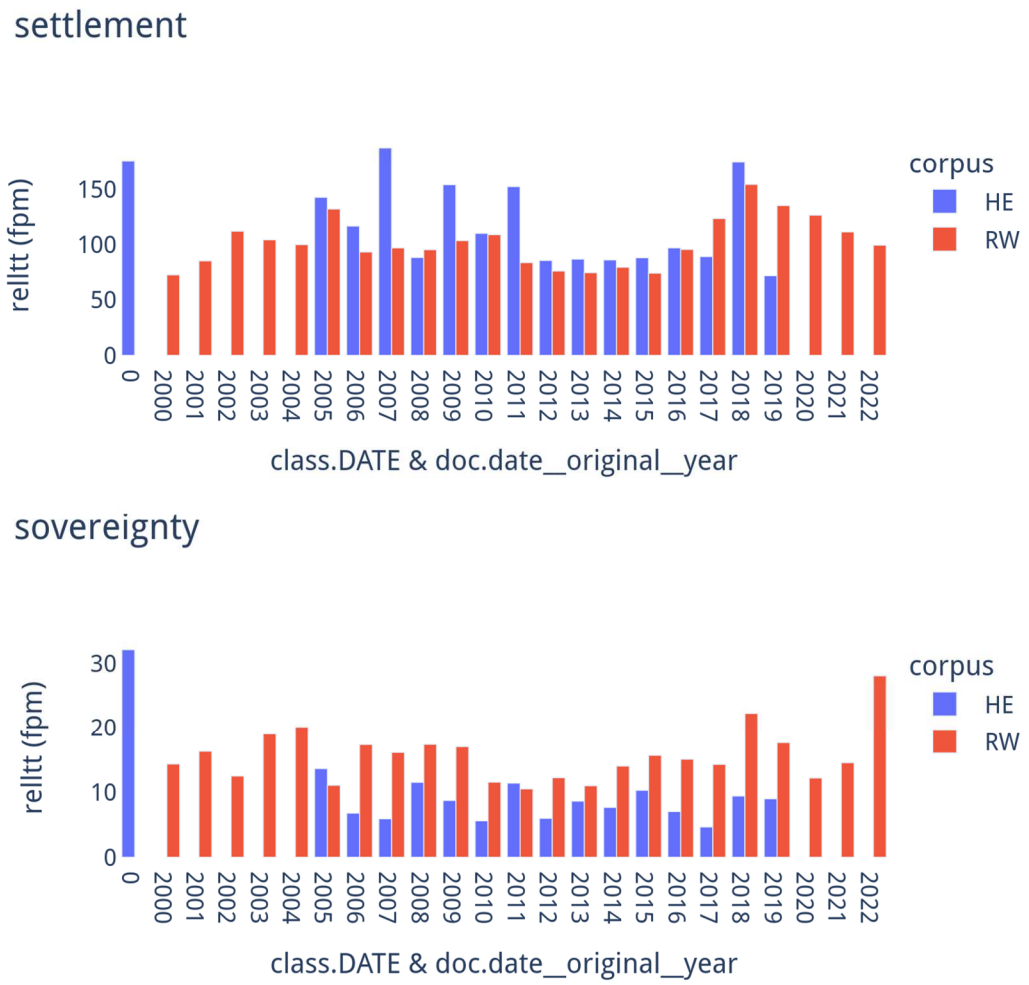

Figure 4: Concepts with K>Q3

Several concepts experienced upward trends over the last two decades, such as INNOVATION, EMPOWERMENT, DO NO HARM, INCLUSION, CASH, CONTEXT, PSYCHOSOCIAL SUPPORT, GENDER-BASED VIOLENCE, RESILIENCE, ARMED ACTORS, and FORCED DISPLACEMENT. Figure 5 shows annual relative frequencies for RESILIENCE and GENDER-BASED VIOLENCE beginning near 0 and reaching close to 400 and 200, respectively, as part of generally steady increases.

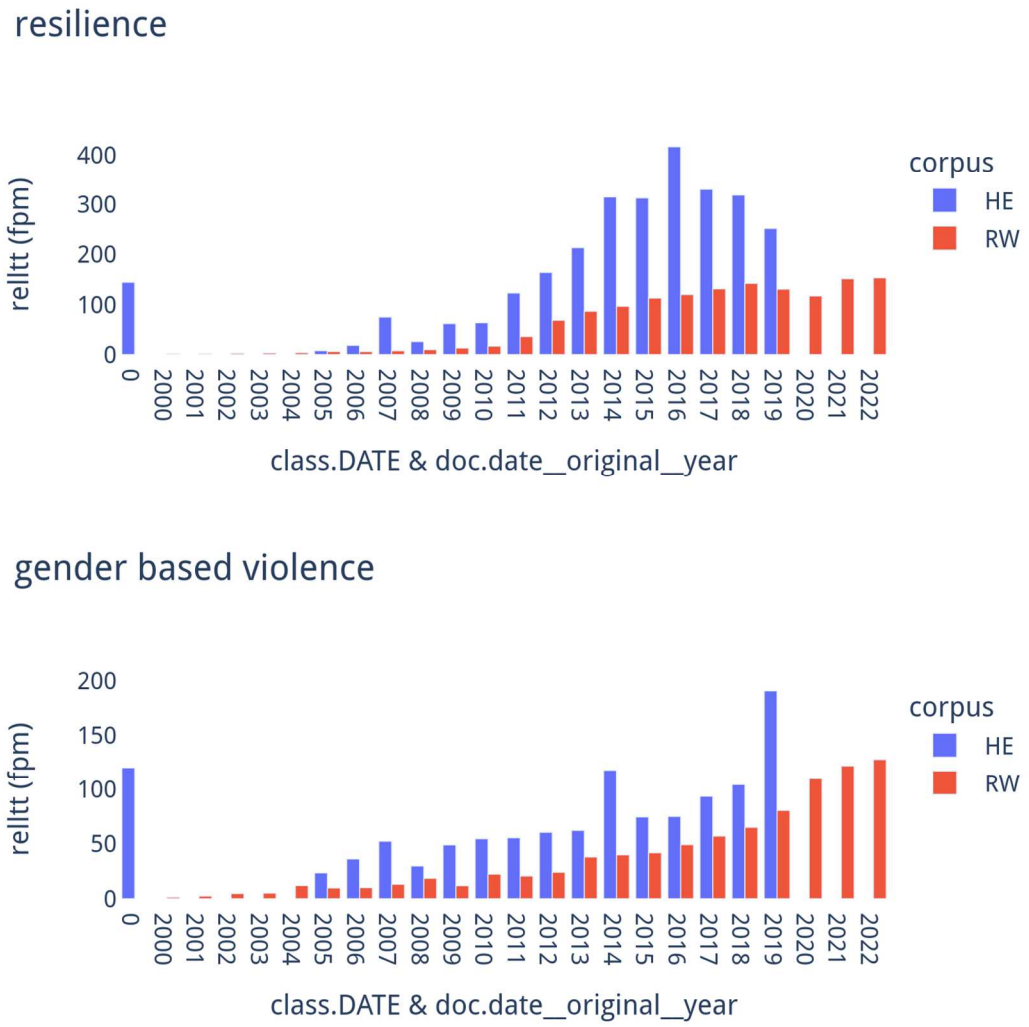

Figure 5: Concepts with shared upward trends

3.3 Hypernym and definitional context comparison

In randomized samples of up to 1,000 contexts (where possible), the average density of hypernyms and definitional contexts fluctuated from 0.10% for SETTLEMENT in RW to 5.40% for GENDER-BASED VIOLENCE in RW. Overall, RW had slightly higher KRC densities for the six concepts, at an average of 2.48% against 2.23% in HE. Two concepts had several-fold differences in density, with KRCs for HUMANITARIAN REFORM being 3.65 times more frequent in RW and KRCs for SETTLEMENT being 12.00 times more frequent in HE. These two concepts happened to be the least and most frequent in absolute terms (509 and 89,283 concordances) in RW.

Concept	*K*	Concordances		KRCs		Density %	
		HE	RW	HE	RW	HE	RW
humanitarian reform	0.143	699	509	3	8	0.43	1.57
sustainability	0.274	9,060	12,614	20	29	2.00	2.90
resilience	0.603	17,789	54,437	13	12	1.30	1.20
gender-based violence	1.135	5,991	34,516	40	54	4.00	5.40
settlement	1.837	9,572	89,283	12	6	1.20	0.10
sovereignty	3.85	701	13,692	31	32	4.42	3.20
mean	1.307	7,302	34,175	19.8	23.5	2.23	2.48

Note: Sample size = 1,000 random concordances or all if fewer

Table 5: Density of hypernymic and definitional contexts

Among the 260 contexts extracted for the six concepts were 104 monolexical hypernyms (including repeated cases, e.g., with ISSUE appearing separately for three concepts). 25 hypernyms, or 24%, were shared for the same concept across corpora, with HE having 34 additional hypernyms and RW 45. Once again, HUMANITARIAN REFORM and SETTLEMENT stood out for having the fewest shared hypernyms (0 of 9 for the former and 1 of 16 for the latter). In contrast, SOVEREIGNTY had the most homogeneous hypernyms, with 4 of 9 being shared (44%).

Concept	Shared	HE	RW
humanitarian reform	(0/9)	challenge, development [recent change], matter	module, initiative, issue, priority, reform, solution
sustainability	criterion, goal, indicator, issue, principle, theme, topic (7/25)	category, cornerstone, driver, objective	area, catchword, challenge, component, concept, concern, element, journey, measure, pillar, point, priority, problem, struggle
resilience	area, capacity, concept, term (4/13)	ability, notion, objective, priority, theme	accelerator, buzzword, pillar, quality
gender-based violence	abuse, challenge, concern, crime, issue, problem, term, violation, violence (9/32)	act, area, burden, component, crisis, practice, precursor, reaction, topic, weapon	barrier, discrimination, epidemic, exploitation, fact, injustice, phenomenon, plague, risk, scourge, threat, trauma, vulnerability
settlement	area (1/16)	bargain, categorization, concern, crime, need, shelter, slum, town	action, activity, community, facility, measure, village, violation
sovereignty	concept, issue, notion, principle (4/9)	priority, responsibility, right, theme	idea

Table 6: Shared and unique monolexical hypernyms

Among the contexts collected from concordance samples, a small minority were definitional, with only three of the concepts having this type of KRC in both corpora. RW had one context for SUSTAINABILITY, albeit less formal: "sustainability entails "striking a balance between the needs of both human and natural systems"". RESILIENCE had four definitional contexts in HE, all similar to the example offered in Table 7. In contrast, each of the five contexts for RESILIENCE in RW were complementary but contextualized the concept in distinct settings: road infrastructure and farming, social institutions, livelihood systems, difficult situations, and cities. GENDER-BASED VIOLENCE had five contexts in HE and four in RW, with one subsumed in a definition for VIOLENCE AGAINST WOMEN. SOVEREIGNTY had four contexts in HE, with three being repeats from the same organization (due to the small sample size), and one in RW. For both corpora, these definitional contexts specifically treated FOOD SOVEREIGNTY rather than SOVEREIGNTY generically.

Concept	HE	RW
resilience	GOAL defines resilience as "the ability of communities and households living within complex systems to anticipate and adapt to risks, and to absorb, respond and recover from shocks and stresses in a timely and effective manner without compromising their long term prospects, ultimately improving their well-being.	Resilience refers here to the capacity of these social institutions to absorb and adapt in order to sustain an acceptable level of functioning, structure, and identity under stress.
gender-based violence	This Strategy defines GBV "as violence that is directed at an individual based on his or her biological sex, gender identity, perceived adherence to socially defined norms of masculinity and femininity.	Gender-based violence (GBV) is an umbrella term for any harmful act perpetrated against a person's will based on the socially ascribed (i.e. gender) differences between females and males.
sovereignty	In its own words: "Food sovereignty is the right of peoples to healthy and culturally appropriate food produced through sustainable methods and their right to define their own food and agriculture systems.	Though closely linked to food insecurity, food sovereignty involves the right of a state to be food self-sufficient based on their own democratically-determined polices.

Table 7: Definitional contexts across corpora

4. Discussion

Despite the shared domain of the corpora, the relative frequencies of humanitarian concepts in ReliefWeb's sometimes abbreviated HTML content are very often lower than the Humanitarian Encyclopedia's complete texts. This is especially the case when comparing per million token frequencies in each year shared by the corpora, which offers a more accurate depiction of how common concepts are. Using keyness scores that adjust for shared years, only 1.55% of concepts were at least twice as common in RW, whereas 67.44% were at least twice as common in HE.

That said, the size and scope of the RW corpus offered data that paralleled the HE corpus at each stage in the analysis. Diachronic trends for both stable and unstable concepts often agreed, the density of hypernymic contexts was similar and had

important overlaps, and both the appearance of and content of definitional contexts generally coincided. In other words, despite the large disparity in the average length of texts in each corpus (a 26-fold difference), analysis results for key humanitarian concepts are likely to share many commonalities.

Definitional contexts were found in both corpora precisely for the two concepts that experienced increasing usage in the previous two decades (RESILIENCE and GENDER-BASED VIOLENCE), whereas comparatively stable concepts like SETTLEMENT had no definitions. The one exception was FOOD SOVEREIGNTY, indicating that definitions for important yet less common hyponyms may also be captured to a similar extent. Still, the frequency of hypernyms varied widely, as with the much-reduced frequency of KRCs for SETTLEMENT in RW. The overall number of overlapping hypernyms was also low; along with the varied definitions for RESILIENCE in RW, it is apparent that RW contributes important diversity for some concepts, regardless of keyness.

While an analysis of concepts by organization type and theme would be beneficial in future work, one can still note that the concepts with the highest keyness in RW tend to underscore ReliefWeb's focus on emergency response. There was a preponderance of EVENT concepts in Q4 that afflict populations (EMERGENCY, GENDER-BASED VIOLENCE, CONFLICT, FORCED DISPLACEMENT, EPIDEMIC, CRIME, FAMINE, TERRORISM, EVACUATION). In contrast, Q1 contained more abstract and process-oriented concepts related to humanitarian action (ADVOCACY, EFFICIENCY, INNOVATION, SUSTAINABILITY, KNOWLEDGE, MANAGEMENT, TECHNOLOGY, DEVELOPMENT, ACCOUNTABILITY). This divergent focus between the corpora may be an important consideration particularly when studying humanitarian development practices with ReliefWeb's curated HTML reports.

Although this analysis offered perspective on ReliefWeb's composition, as well as some characteristics relevant to the study of humanitarian concepts, a main limitation was its restriction to HTML content. Including PDF content would provide a more complete vision, likely increasing the relative frequencies of the domain's core concepts. This task, which is underway, requires a more advanced pipeline with text extraction and language identification. Still, the data collected validate that KRC-based concept analysis can be fruitful with the HTML texts. While a workflow was developed here to build and update ReliefWeb corpora, in English and other languages, optimizing data extraction and its presentation to the humanitarian community is another area to contend with.

5. Conclusion

This paper presented a corpus of two thirds of HTML reports available on the United Nations-managed service ReliefWeb. These were mostly short news articles, press releases, or summaries in English regarding humanitarian response to emergency events. The corpus was compiled and inspected with a mix of open-source software,

including the Stanza NLP package, NoSketch Engine, and a Python package (Corpusama, available on GitHub) that was introduced to manage corpus generation with the service's API.

A keyness analysis comparing 129 humanitarian concepts in the ReliefWeb corpus with a corpus developed by the Humanitarian Encyclopedia showed that ReliefWeb's HTML content has consistently low relative frequencies for these concepts. Still, a subsequent knowledge-rich context analysis of six concepts (HUMANITARIAN REFORM, SUSTAINABILITY, RESILIENCE, GENDER-BASED VIOLENCE, SETTLEMENT, SOVEREIGNTY) indicated that the corpus offers both similar and complementary data for hypernym- and definition-centered information extraction. This result contextualizes the database's potential and limits for humanitarian concept analysis, including the sort conducted by the Humanitarian Encyclopedia.

In the future, the development of a multilingual family of ReliefWeb corpora compatible with popular language corpus management software could be a boon for studying humanitarian discourse across linguistic communities. This is a goal for future versions of the Corpusama package. The current analysis, which focused on the composition of the English corpus and the frequencies of key humanitarian concepts, did not take into account most of the metadata offered via the service's API. Organization type, document format, and humanitarian theme are prime candidates for more research, in the form of larger, more complete concept analyses and further inspection of the various characteristics of humanitarian reports on ReliefWeb.

6. Acknowledgments

Funding for this work was provided through the Humanitarian Encyclopedia project at the Geneva Centre of Humanitarian Studies and the research project PROYEXCEL_00369 (VariTermiHum), funded by the Regional Government of Andalusia (Spain).

7. References

Chambó, S., & León-Araúz, P. (2021). Visualising lexical data for a corpus-driven encyclopaedia. In I. Kosem, M. Cukr, J. Miloš, J. Kallas, S. Krek, & C. Tiberius (eds.) *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 Conference.* Brno, Czech Republic: Lexical Computing, pp. 29–55.

Condamines, A. (2022). How the notion of "knowledge rich context" can be characterized today. *Frontiers in Communication*, 7. https://doi.org/10.3389/fcomm.2022.824711

Faber, P. (2022). Frame-based terminology. In P. Faber & M.-C. L'Homme (eds.) *Theoretical Perspectives on Terminology.* Amsterdam: John Benjamins, pp. 353–376.

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (eds.) *Corpus Approaches to Discourse: A Critical Review.* London: Routledge, pp. 225–258.

Isaacs, L. (2022). Sketch Grammar Explorer (Version 0.5.5) [Computer software]. https://doi.org/10.5281/zenodo.6812335

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. In R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, & Y. Harada (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 2010).* Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp. 741–747.

Kilgarriff, A. (2012). Getting to know your corpus. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (eds.) *Text, Speech and Dialogue 15th International Conference, TSD 2012.* Brno, Czech Republic: Springer, pp. 3–15. https://doi.org/10.1007/978-3-642-32790-2_1

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), pp. 7–36. https://doi.org/10.1007/s40607-014-0009-9

León-Araúz, P., & San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: From knowledge patterns to word sketches. In I. Kerneman & S. Krek (eds.) *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets".* Miyazaki, Japan: Globalex, pp. 94–99.

León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin, A. Francoeur, J. Humbley, & A. Picton (eds.), *Multiple Perspectives on Terminological Variation.* Amsterdam: John Benjamins, pp. 213–258.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), pp. 313–330.

Marshman, E. (2022). Knowledge patterns in corpora. In P. Faber & M.-C. L'Homme (eds.) *Theoretical Perspectives on Terminology.* Amsterdam: John Benjamins, pp. 291–310.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (eds.) *Recent Advances in Computational Terminology (Vol. 2).* Amsterdam: John Benjamins, pp. 279–302. https://doi.org/10.1075/nlp.2.15mey

Naidoo, S. (2007). Redesigning the ReliefWeb. *Information Management Journal*, 41(5), pp. 52–58.

National Laboratory for Digital Heritage, Eötvös Loránd University Department of Digital Humanities. (2023). NoSketch-Engine-Docker (Version 5.0.0) [Computer software]. https://github.com/ELTE-DH/NoSketch-Engine-Docker

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python

natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 101–108.

Rubin, O. (2014). Diagnosis of famine: A discursive contribution. *Disasters*, 38(1), pp. 1–21. https://doi.org/10.1111/disa.12030

Ruso, S. (1996). ReliefWeb: Mandate and objectives. *Refuge*, 15(4), pp. 18–20. https://doi.org/10.25071/1920-7336.21881

Rychlý, P. (2007). Manatee/Bonito — a modular corpus manager. In P. Sojka & A. Horák (eds.) *First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007*. Brno, Czech Republic: Masaryk University, pp. 65–70.

San Martín, A., Trekker, C., & León-Araúz, P. (2020). Extraction of hyponymic relations in French with knowledge-pattern-based word sketches. In N. Calzolari et al. (eds.) *Proceedings of The 12th Language Resources and Evaluation Conference (LREC-2020)*. Marseille, France: European Language Resources Association, pp. 5953–5961.

Shamoug, A., Cranefield, S., & Dick, G. (2023). SEmHuS: A semantically embedded humanitarian space. *Journal of International Humanitarian Action*, 8(3). https://doi.org/10.1186/s41018-023-00135-4

Sierra, G., Alarcón, R., Aguilar, C., & Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1), pp. 74–98. https://doi.org/10.1075/term.14.1.05sie

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association, pp. 2897–2904.

United Nations Office for the Coordination of Humanitarian Affairs. (2022). ReliefWeb analytics: 2022 mid-year highlights. https://reliefweb.int/report/world/reliefweb-highlights-mid-year-2022

Von Schreeb, J., Legha, J. K., Karlsson, N., & Garfield, R. (2013). Information for action? Analysis of 2005 South Asian earthquake reports posted on Reliefweb. *Disaster Medicine and Public Health Preparedness*, 7(3), pp. 251–256. https://doi.org/10.1001/dmp.2010.36

Wackernagel, M., & Footner, A. (2021, October 6). Talking Heads: ReliefWeb then and now. *ReliefWeb*. https://reliefweb.int/blogpost/talking-heads-reliefweb-then-and-now