Humanitarian Corpora for English, French and Spanish

Loryn Isaacs, Santiago Chambó, Pilar León-Araúz
Department of Translation and Interpreting, University of Granada
Puentezuelas, 55, 18071, Granada, Spain
{lisaacs, santiagochambo, pleon}@ugr.es

Abstract

This paper presents three corpora of English, French and Spanish humanitarian documents compiled with reports obtained from ReliefWeb through its API. ReliefWeb is a leading database of humanitarian documents operated by the UN Office for the Coordination of Humanitarian Affairs (OCHA). To compile these corpora, documents were selected with language identification and noise reduction techniques. They were subsequently tokenized, lemmatized, tagged by part of speech, and enriched with metadata for use by linguists in corpus query software. These corpora were compiled to satisfy the research needs of the Humanitarian Encyclopedia, a project with a focus on conceptual variation. However, they can also be useful for other humanitarian endeavors, whether they are research- or practitioner-oriented; the source code for generating the corpora is available on GitHub. To compare materials, an exploratory analysis of definitional and generic-specific information was conducted for the concept of ARMED ACTOR with lexical data extracted from an English legacy corpus (where the concept is underrepresented) as well as on the new English and Spanish corpora. Lexical data were compared among corpora and presented by means of online data visualization to illustrate its potential to inform conceptual modelling.

Keywords: corpus creation, ReliefWeb, Humanitarian Encyclopedia, conceptual variation

1. Introduction

ReliefWeb is a service operated by the UN Office for the Coordination of Humanitarian Affairs (OCHA) that archives publicly available humanitarian documents. It has positioned itself as the main information system for humanitarians, with a database containing over a million reports. ReliefWeb also offers access to all its content through a publicly accessible API that takes HTTP requests and returns JSON data. ReliefWeb's API[1] is well-documented and accessible to third-party developers.

This paper presents three corpora of English, French and Spanish humanitarian documents compiled with reports obtained from ReliefWeb through its API. These corpora were tokenized, lemmatized, and tagged by part of speech to make them compatible with corpus management and querying software. They also contain rich corpus metadata, making them a valuable resource for research in multilingual humanitarian communication. This article is structured as follows: Section 2 presents the context and needs that motivated the compilation. Section 3 details the materials and methods employed, as well as the composition of the corpora. Section 4 presents a preliminary analysis of lexical data and discusses its usefulness for the purposes of corpus-driven conceptual analysis. Finally, in Section 5, a conclusion is drawn and future uses of the corpora are outlined.

2. Background

The humanitarian domain is a recent professional field (Eberwein and Saurugger, 2013), which encompasses around 5,000 organizations (ALNAP, 2022) with diverse specializations, organizational cultures and perspectives on humanitarian action (Dauvin and Siméant-Germanos, 2002; Sezgin and Dijkzeul, 2015). Humanitarian actors and scholars have generated critical self-reflection and competing conceptualizations of the domain (Rey Marcos, 2003), as well as shared terminologies (Reliefweb, 2008). However, humanitarians often report on the need of shared understandings of key concepts to improve stakeholder coordination (OCHA, 2012, p. 2), produce common operationalizations for comparable measurement of phenomena, and better guide decision-making (IFRC, 2016, pp. 199-201).

Fuzzy understandings of humanitarian concepts may lead to complex operational consequences. An example of this is the concept of LOCAL ORGANIZATION. In 2016, leading humanitarian organizations agreed to strengthen the capacity of local organizations to respond to crises (WHS, 2016, p. 36). This came to be known as the localization agenda. Recently, researchers have demonstrated that different understandings of LOCAL ORGANIZATION have resulted in disparate implementations of the agenda, which have paradoxically reduced the capacity of local actors to engage in humanitarian response (Khan and Kontinen, 2022; Roepstorff, 2020). Conceptual variation (see Hampton, 2020) hampers policy implementation and, therefore, has the potential to reduce the effectiveness of humanitarian action.

In 2018, the Humanitarian Encyclopedia (HE; humanitarianencyclopedia.org) was launched as a collaborative initiative to describe 129 ill-defined concepts with a focus on conceptual variation. The HE is an ongoing project, was initiated by the Geneva Centre of Humanitarian Studies (humanitarianstudies.ch) and is currently managed by a newly-formed consortium of humanitarian organizations and academic institutions. It aims to foster a common understanding of humanitarian concepts and improve collective humanitarian action by becoming an influential descriptive reference work

[1] https://apidoc.rwlabs.org/

LREC-COLING 2024, pages 8418–8426

for humanitarians (Humanitarian Encyclopedia, 2021c).

The HE employs knowledge transfer (Ward et al., 2012) between linguists and humanitarian experts to describe concepts. Entries are written by experts who are provided with corpus-based conceptual analysis reports produced by linguists. This serves two purposes. Firstly, it aims to reduce possible biases and content gaps due to the diverse professional and academic backgrounds of entry authors. Secondly, it informs experts of potential cases of conceptual variation by obtaining lexical data from a wide range of textual sources (Humanitarian Encyclopedia, 2021b).

Linguists analyze concepts with a methodology derived from Frame-based Terminology (Faber, 2022). Textual evidence of conceptual description is obtained by extracting definitions, knowledge-rich contexts (KRCs) (Marshman, 2022), and multi-word term expressions (MWT) (Cabezas-García and Gil-Berrozpe, 2018) from corpora with semantic pattern-based extraction techniques (León-Araúz and San Martín, 2018; San Martín et al., 2020). Conceptual characteristics are elucidated by modelling lexical data from KRCs into conceptual propositions (i.e., pairs of concepts linked by semantic relations) and associating them with corpus metadata to operationalize conceptual variation (Chambó and León-Araúz, 2023). Results are presented in reports with definitional templates (Durán-Muñoz, 2016; León-Araúz et al., 2012) and interactive data visualizations.

To date, linguists have used the HE Corpus (Egger et al., 2018; Humanitarian Encyclopedia 2021a, which was complied to serve as the basis for conceptual analyses. It contains 4,824 humanitarian English documents published between 2004 and 2019, amounting to nearly 85 million tokens. Corpus metadata include an *a priori* taxonomy of organization types and subtypes (e.g. NGOs, IGOs, etc.), region of publication, year of publication and document typology (general document, strategy document and activity report).

Currently, the HE needs a larger corpus for various reasons. Firstly, conceptual analysis work is still ongoing, which has rendered the HE Corpus obsolete, as it lacks documents for the 2020-2023 period. Secondly, the HE Corpus lacks sufficient frequency counts for several concepts. In fact, it contains less than 1,000 occurrences for 19 of the 129 concepts and less than 5,000 occurrences for 46 concepts. Only 8 concepts are represented with over 10,000 occurrences in the corpus. While low-frequency conceptual analyses have been performed, data were insufficient to draw meaningful comparisons among subcorpora for any valid study of conceptual variation. Thirdly, a new research project has extended the scope of study to Spanish humanitarian discourse. This will require great amounts of lexical data to identify key concepts specific to the Spanish linguistic community, perform conceptual analysis and study interlinguistic conceptual variation.

To expand the HE Corpus, ReliefWeb was an easy choice given its prominent status, the great number of documents it contains and its friendliness to third-party developers. A key aspect is also the corpus metadata provided, given that it is an essential requirement to study conceptual variation. Another benefit is that ReliefWeb continuously incorporates new documents, making future updates of corpora feasible. A preliminary corpus of ReliefWeb's HTML content (mainly short news and press release items) was previously compiled to perform a comparative assessment against the HE Corpus (Isaacs, 2023b). This demonstrated the suitability of ReliefWeb as a humanitarian corpus by analyzing concept keyness and diachronic trends, as well as the density of hypernymic and definitional KRCs for a sample of concepts. The next section builds on this previous work and describes the compilation process to include full-length PDF content from ReliefWeb.

3. Corpus compilation

ReliefWeb's API was utilized in conjunction with the Corpusama tool (Isaacs, 2023a) to generate English, French and Spanish corpora. These included reports from 2000 through June 2023, including HTML and PDF content. As described in this section, they represent about 95% of ReliefWeb data for the languages during this time span and amount to over 2.36 billion tokens. The workflow builds off previous efforts; novel features are summarized below.

Large-scale manipulation of ReliefWeb's database has been undertaken on several occasions in natural language processing (NLP) literature. To our knowledge, one dataset has been made publicly accessible (Horwood, 2017) and utilized on several occasions (Littell et al., 2018; Muis et al., 2018; Shamoug et al., 2023). Other research has also extracted large portions of data (Nemeskey and Kornai, 2018; Papadopoulos et al., 2017; Tamagnone et al., 2023). The Corpusama tool was developed in part to establish a fully reproducible and updatable means to generate corpora from ReliefWeb: see the project's GitHub page for source code and updates to corpus availability.[2] Hence, the corpora introduced in the current work represent an ongoing process for improving the accessibility and quality of the data for NLP tasks generally but corpus linguistics more specifically.

3.1 Language Identification

Existing text layers of PDFs associated with ReliefWeb reports were extracted with PyMuPDF, which was preferred over three other tools: pypdf, pdfminer.six, and PDFBox. These were assessed based on availability, active development, ease of implementation, speed and previous comparisons

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

(Bast and Korzen, 2017; Miah et al., 2022). Documents were then normalized with uninorm_4, a script used by Sketch Engine[3] and part of UniTok(Michelfeit et al., 2014). Further cleaning was applied temporarily to minimize noise that could interfere with language identification (LI), including symbol and punctuation removal and all-uppercase to lowercase line conversion (where a line is a string of characters ending with a control character, e.g., \n).

LI was conducted with fastText (Joulin et al., 2016). This was selected among LI tools capable of sorting texts with an unknown combination of languages. A primary factor was its speed on devices lacking a GPU, having performed over 60 times faster than the model used previously for English ReliefWeb content (Qi et al., 2020).

Each line in a file was processed with fastText to guess languages and assign probability scores. Lines with a confidence lower than 0.6 were labeled "unknown" and those with 10 characters or fewer were labeled "short". The size in bytes was calculated for each language in a file and those with insufficient representation were ignored: this method was modelled off another for calculating multilinguality in texts (Abadji et al., 2022). For each document, the LI process produced a dictionary with languages and proportions: for example, one trilingual document obtained {"fr": 0.37, "es": 0.36, "en": 0.26}.

3.1.1 ReliefWeb Database Composition

A total of 1,101,560 HTML (819,877) and PDF (281,683) documents were processed from ReliefWeb's database. Excluded were 6,032 with no text content or extraction errors. Almost 97% were considered monolingual, i.e., had an LI dictionary containing only one language. As shown in Table 1, English made up the great majority, followed by French, Spanish and Arabic. Of the 33,533 other documents, thousands may qualify as monolingual in EN, FR, ES or AR, although with a notable amount of other content that could contribute noise.

Language	Files	Percentage
English	874,757	79.41%
French	106,946	9.71%
Spanish	79,278	7.20%
Arabic	7,046	0.64%
Other*	33,533	3.04%
Total	*1,101,560*	*100%*

*Includes "unknown", "short" and combinations

Table 1: Language identification results

A large portion of "other" documents experienced LI challenges, with 5,575 having at least 50% "unknown" content and 1,668 "short" content. Samples of 20 documents for each of these categories suggest that lists, tables, short phrases, uncommon languages (Tetum, Tongan, Somali) and PDF extraction are factors. Beyond encoding issues and absent text layers, improperly extracted PDFs were occasionally segmented into one word per line.

Some 2,214 documents were identified with an English and French pairing, making this the most common bilingual combination. However, a large portion are likely monolingual French documents with English headings, tables and so forth. Multilingual content valuable for other research purposes (e.g., developing parallel corpora of translations) is therefore expected to be much lower. Additional languages with 1,000 to 2,000 monolingual documents were Sinhala, Ukrainian and Russian, with various others having hundreds, including Farsi, Sorani, Portuguese and Chinese.

3.1.2 Noise reduction

After generating LI results, datasets for creating monolingual corpora were selected via a minimum content threshold. Ultimately, having at least 80% of a document's content (in bytes) in a desired language was required. This was chosen after assessing the noisiness of results at different intervals, from 40% to 95%. 95% was considered restrictive, excluding valuable documents with tables containing numeric data or short, hard-to-identify strings. Conversely, 60% allowed an excessive number of multilingual texts. The final datasets had 858,657 documents for English, 104,602 for French and 76,919 for Spanish. The 80% threshold reduced corpus sizes by close to 5% (3.57% EN, 5.26% FR and 4.97% ES). One PDF, for report no. 1099416, just over the minimum threshold, exemplifies the challenge of reducing multilingual noise while including valuable content.[4]

3.2 Pipelines

Documents included in each corpus were given XML tags with standardized text type metadata and combined into text files that could be passed to FreeLing (Padró and Stanilovsky, 2012) corpus creation pipelines. Sketch Engine's open-source French (v3) and Spanish (v6) pipelines were utilized without substantive modification, and an English pipeline was developed with a similar configuration. This simplified tokenization, lemmatization, and part-of-speech tagging by relying on one NLP package, but also differs from the TreeTagger pipeline common among Sketch Engine's English corpora and the Stanza pipeline used previously. Another consideration was the time and computing power that would have been required to utilize novel machine learning models. These concerns took priority over cutting-edge pipeline accuracy; this aspect of the methodology could be evaluated and improved upon in future work.

[3] https://www.sketchengine.eu

[4] https://reliefweb.int/node/1099416

3.3 Corpus Composition

The corpora were compiled locally with NoSketch Engine (Kilgarriff et al., 2014; Rychlỳ, 2007). Subcorpora for HTML and PDF content were created by adding a file_id tag of 0 to HTML documents and sorting the content (PDFs on ReliefWeb already contain unique non-zero file_id tags). The corpora amounted to 2.36 billion tokens from over 1 million files, with PDFs making up three quarters of tokens.

Corpus	Docs	Types M	Tokens M	PDF tokens
EN	858,657	1,608	1,983	78.25%
FR	104,602	196	235	74.26%
ES	76,919	118	142	77.80%
Total	*1,040,178*	*1,922*	*2,360*	*76.77%**

*Mean

Table 2: Corpora sizes

3.3.1 Text type distributions

The 19 text types taken from API metadata include country of interest, disaster type, document format (genre), report title, source organization, source organization type, humanitarian theme and date. Several are highlighted below (with measurements in number of tokens, unless stated otherwise).

The most-covered geographic area for all three corpora is World. Top individual countries differ substantially: RW_EN is spread across Africa, Eastern Mediterranean and South-East Asia; besides Haiti and Syria, RW_FR focuses on Africa; RW_ES focuses on the Americas, excepting DR Congo, Afghanistan and Myanmar. International Organization is by far the largest source type,[5] from 46-54% across the corpora. Non-governmental Organization and Government are also large contributors. Themes, while more diverse, include Protection and Human Rights, Health, Food and Nutrition, and Water Sanitation Hygiene (WASH). The most common formats by far are Analysis, News and Press Release, and Situation Report. Table 3 and Figures 1-3 show data for several key text types.

Corpus	Country	Source type	Theme
EN	World Afghanistan Syria Sudan Somalia	Intl. Org. NGO Academic Govt. Red Cross	Protection Health Food WASH Education
FR	World DR Congo Mali CAR Niger	Intl. Org. NGO Govt. 2 Intl. Orgs. Academic	Protection Health Food Agriculture WASH
ES	World Colombia Venezuela Peru Guatemala	Intl. Org Govt. NGO Academic 2 Intl. Orgs.	Protection Health Food Education WASH

Table 3: Top text type values by tokens

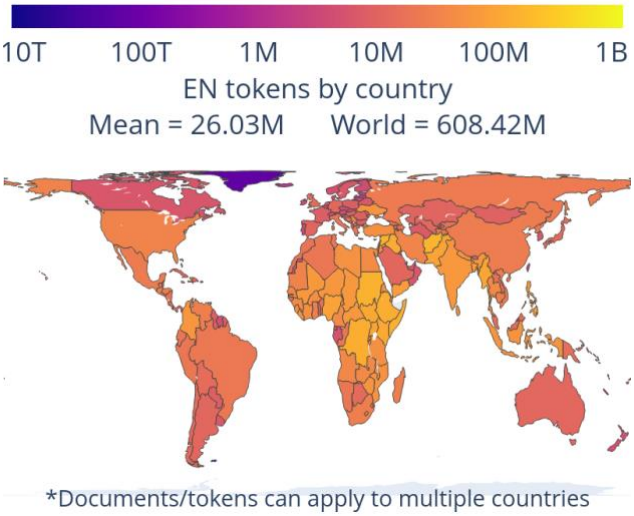

Figure 1: English corpus geographic coverage

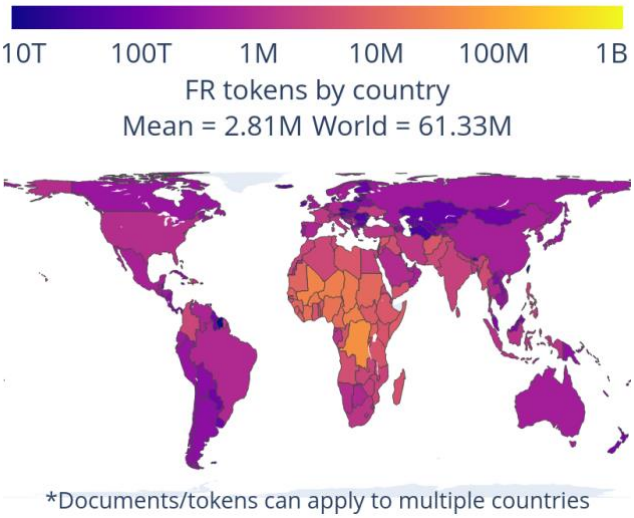

Figure 2: French corpus geographic coverage

[5] https://api.reliefweb.int/v1/references/organization-types

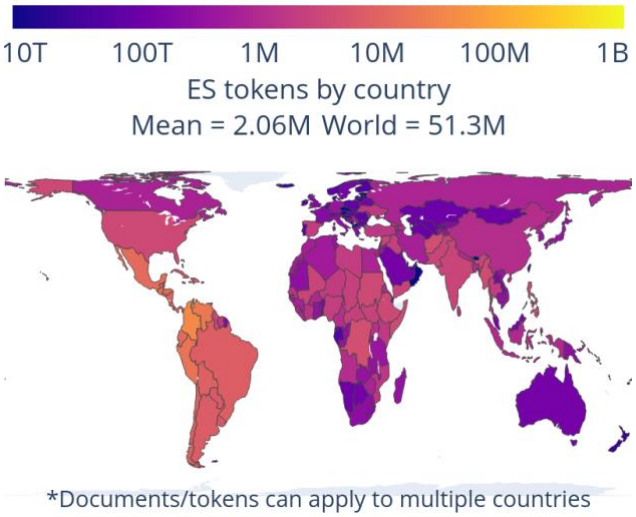

Figure 3: Spanish corpus geographic coverage

4. Probing the corpus

An exploratory analysis of lexical data was conducted to illustrate the potential of the new corpora for the purposes of the HE. The low-frequency concept of ARMED ACTOR (299 occurrences in the HE Corpus vs 15,141 in ReliefWeb EN) was chosen because no conceptual analysis was performed previously given its minimal presence in the original corpus. Other low-frequency concepts among the selected 129 include DO NO HARM (365 vs 12,517), HUMANITARIANISM (355 vs 7,011) and RESPONSIBILITY-TO-PROTECT (317 vs 16,058). This section presents the results of an initial contrastive analysis of definitions and lexical unit candidates for hyponymic and hypernymic modelling, as the identification of dynamic categorizations is the starting point of any analysis addressing conceptual variation. These were extracted from the HE Corpus and the ReliefWeb EN and ES corpora presented here. Corpora were queried in NoSketch Engine with pattern-based knowledge extraction strategies (Isaacs et al., 2024).

4.1 Definition extraction for ARMED ACTOR

Definitions were targeted in a manner similar to Kovář, Močiariková, and Rychlý (2016) by means of definitional verbal and paralinguistic patterns (Sierra et al. 2010; Dorantes et al. 2017). A total of 8 definitions were found in ReliefWeb EN, whereas none was obtained from HE Corpus or ReliefWeb ES, except for a statement that armed groups in Colombia do not fit in conventional definitions. In 7 definitions, *definienda* are MWT hyponyms of ARMED ACTOR, namely 'non-state armed actor' and 'ethnic armed actor'. An example definition for each kind of *definiendum* is provided in Table 4.

(1) armed actor	From 1 January to 30 September, the Gender-Based Violence Information Management System (GBVIMS) recorded 30 incidents perpetrated by armed actors (defined by the GBVIMS as members of an armed group, police, State military or security officials). (OHCHR, 2021)
(2) non-state armed actor	A definition of non-state armed actors has proven difficult owing to their many types and characteristics. Generally speaking, non-state armed groups are defined as distinctive organizations that are (i) willing and capable to use violence for pursuing their objectives and (ii) not integrated into formalized state institutions such as regular armies, presidential guards, police, or special forces. They, therefore, (iii) possess a certain degree of autonomy with regard to politics, military operations, resources, and infrastructure. They may, however, be supported or instrumentalized by state actors either secretly or openly, as happens often with militias, paramilitaries, mercenaries, or private military companies. (Hofmann and Schneckener, 2011)
(3) ethnic armed actor	Ethnic Armed Actor: a broad term referring to all armed actors explicitly associated with ethnic nationalities, including: state-backed militia, border guard forces, opposition groups that maintain ceasefires, and those that are actively fighting with government forces. (Jolliffe, 2015)

Table 4: Definitions for ARMED ACTOR

The most detailed definitions are those for NON-STATE ARMED ACTOR, as illustrated by (2) in Table 4. In definitions for NON-STATE ARMED ACTOR and ETHNIC ARMED ACTOR, *definienda* are described in terms of collective entities designating organized human groups. However, as can be seen in (1), ARMED ACTOR is defined in the context of a gender-based violence reporting system, for which instances of the concept designate individual people who belong to armed organizations and are reported as perpetrators of violence. NON-STATE ARMED ACTOR appears to be the concept that attracts most definitional effort, which may indicate both the relevance of the concept and its ill-defined nature. Furthermore, as shall be seen in the next section, this specific type of armed actor is the one most frequently highlighted across all organizations in the ReliefWeb EN corpus.

4.2 Hypernyms and hyponyms of ARMED ACTOR

Two strategies were used to extract lexical candidates that may inform hypernymic and hyponymic modelling. Firstly, KRCs were exhaustively extracted with generic-specific pattern-based techniques (León-Araúz and San Martín, 2018; San Martín et al., 2020) and manually curated. Secondly, MWT hyponyms were extracted by querying the corpora for modifiers before 'armed actor' and their Spanish equivalents. Analyzing catalogues of MWTs is an essential step in conceptual analysis, as many semantic relations can be inferred from their internal structure (Cabezas-García and León-Araúz 2018).

Total counts of KRCs and candidates are detailed in Table 5. As is often the case, while some candidates stand out with relatively high frequencies, the bulk of candidates display high heterogeneity with low individual counts. To illustrate this, interactive

visualizations were created with flourish.studio to display all candidates disaggregated by organization type as per corpus metadata; links are provided in Table 6. Only three hypernymic candidates were obtained from ReliefWeb ES. These were not visualized but are discussed in Section 4.3.

Corpus	KRCs	Hypernymic candidates	Hyponymic candidates
HE Corpus	29	9 (7 distinct)	37 (25)
ReliefWeb EN	700	153 (59)	1448 (504)
ReliefWeb ES	212	8 (3)	357 (99)

Table 5: KRCs and candidates from each corpus

Most hyponymic candidates were found for ARMED ACTOR proper. Nonetheless, KRC extraction found candidates for a small group of 34 other MWT hyponyms of ARMED ACTOR such as 'quasi-state armed actor', 'illegal armed actor' and 'non-traditional armed actor', among others, in ReliefWeb EN; 7 in ReliefWeb ES, such as '*actor armado ilegal'* and '*actor con capacidad armada'*; and 1 in the HE Corpus, namely 'non-state actor'. This information is useful for conceptual analysts to develop subcategories.

Corpus	Candidates	Visualization
HE Corpus	Hyponyms	https://public.flourish.studio/visualisation/15435413/
HE Corpus	Hypernyms	https://public.flourish.studio/visualisation/15438966/
HE Corpus	MWT hyponyms	https://public.flourish.studio/visualisation/15434284/
ReliefWeb EN	Hyponyms	https://public.flourish.studio/visualisation/15437412/
ReliefWeb EN	Hypernyms	https://public.flourish.studio/visualisation/15433955/
ReliefWeb EN	MWT hyponyms	https://public.flourish.studio/visualisation/15433520/
ReliefWeb ES	Hyponyms	https://public.flourish.studio/visualisation/15438038/
ReliefWeb ES	MWT hyponyms	https://public.flourish.studio/visualisation/15438114/

Table 6: KRCs and candidates from each corpus

Hypernymic and hyponymic candidates obtained from the new corpora are very diverse and will require extensive inductive categorization. Additionally, many named entities were obtained especially from ReliefWeb EN, which can be used to highlight conflicts that might otherwise be ignored by entry authors in the research process.

4.3 Differences in lexical data between ReliefWeb EN and ES

In ReliefWeb ES, hypernymic candidates are not as abundant. In fact, '*actores armados'* is not explicitly linked to hypernyms, save for three incidental times: (1) in a rather obvious definitional statement (*actores enfrentados que participan en conflictos armados* [confronted actors participating in armed conflicts]); and (2) in highly critical metaphorical accounts on the Colombian conflict: *máquinas de guerra* [war machines] and *cazadores de rentas* [income hunters].

The Colombian conflict appears to drive the most widespread conceptualization of ARMED GROUP in the Spanish humanitarian discourse. Not surprisingly, the most frequent named entities are not ISIS or Al-Shabaab (also present), but FARC, ELN, BACRIM, AUC and AGC.

As for the collective hyponyms extracted, ReliefWeb ES coincides with ReliefWeb EN in that military forces are the subtype that occurs most frequently in international organizations. In contrast, while English academic documents focus on rebel groups and Israeli armed forces, Spanish documents focus on the Colombian military forces. Both government and NGO documents highlight FARC, ELN, guerrillas, and paramilitaries.

For MWT hyponyms, in contrast to ReliefWeb EN, the most frequent type is '*actor armado ilegal'* [illegal armed actor] in nearly all organization types. Only Red Cross and academic documents show more counts for '*actor armado no estatal'* [non-state armed actor], the most common and explicitly defined type in ReliefWeb EN.

However, a definition extracted from ReliefWeb EN sheds light on the matter: "In Colombia, the many non-state armed groups are mostly referred to as illegal armed actors" (Rüttinger et al., 2022: 24). That the terms refer to the same concept is thus well-represented in both English and Spanish discourse. Yet, the preference of '*actor armado ilegal'* over the term variant '*actor armado no estatal'* highlights the illegal status of entities over autonomy from State structures, revealing a case of variation in contrast with English humanitarian discourse.

5. Conclusions

The Humanitarian Encyclopedia (HE) needs larger, up-to-date and multilingual corpora of specialized humanitarian documents to analyze key humanitarian concepts for its entries and study conceptual variation. Three new humanitarian corpora were compiled with HTML and PDF content in English, Spanish and French from ReliefWeb, a leading UN-managed database of humanitarian documents, through its publicly available API. Language identification and noise reduction techniques were employed to select suitable documents, which were subsequently tokenized, lemmatized, and tagged for part of speech for use by linguists in corpus management and query software, as well as enriched with metadata obtained from the API. This method is fully replicable and allows corpora to be updated in the future.

In addition, an exploratory analysis of lexical data was performed for a humanitarian concept that was underrepresented in an English corpus used by the

HE thus far. The analysis was conducted with both English and Spanish lexical data. As expected, the large size of the English corpus provides sufficient data for linguists to conduct analyses that were not previously possible. However, despite the large amount of new data obtained, the substantial difference in corpus size is reflected in the lesser availability of Spanish textual evidence for comparable studies of conceptual variation. Nonetheless, our preliminary analysis suggests that the Spanish conceptualization of ARMED ACTOR may be highly influenced by the Colombian armed conflict.

The size of the new Spanish corpus is greater than any of the materials used by the HE in the past. This enables research into multilingual conceptual variation as well as in other subfields of humanitarian studies. Future efforts will focus on enriching corpora from sources beyond ReliefWeb, especially for languages other than English. This will help offset the predominant international and inter-governmental perspective on the humanitarian domain by collecting a more representative sample of humanitarian discourse.

6. Acknowledgments

This research was funded by the Regional Government of Andalusia (Spain) as part of project PROYEXCEL_00369 (VariTermiHum).

7. Bibliographical References

Abadji, J., Ortiz Suarez, P., Romary, L., and Sagot, B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, 4344–4355, Marseille, France, June. European Language Resources Association (ELRA).

ALNAP. (2022). *The State of the Humanitarian System*. London, United Kingdom. ALNAP/ODI.

Bast, H., and Korzen, C. (2017). A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–10, Toronto, Canada, June. Institute of Electrical and Electronic Engineers (IEEE). https://doi.org/10.1109/JCDL.2017.7991564

Cabezas-García, M., and Gil-Berrozpe, J.C. (2018). Semantic-based retrieval of complex nominals in terminographic resources. In *Proceedings of the XVIII EURALEX International Congress*, 269–281. Ljubljana, Slovenia, July. Ljubljana University Press, Faculty of Arts.

Cabezas-García, M., and León-Araúz, P. (2018). Towards the inference of semantic relations in complex nominals: A pilot study. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2511–2518, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chambó, S., and León-Araúz, P. (2023). Operationalising and representing conceptual variation for a corpus-driven encyclopaedia. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, 587–612, Brno, Czech Republic, June. Lexical Computing.

Dauvin, P., and Siméant-Germanos, J. (2002). *Le travail humanitaire*. Paris, France: Presses de Sciences Po.

Dorantes, M.A., Pimentel A., Sierra G., Bel-Enguix G., and Molina C. (2017). Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos. *Linguamática* 9(2): 33–44. https://doi.org/10.21814/lm.9.2.257

Durán-Muñoz, I. (2016). Producing frame-based definitions: A case study. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(2): 223–249. https://doi.org/10.1075/term.22.2.04mun

Eberwein, W-D., and Saurugger S. (2013). The professionalization of international non-governmental organizations. In B. Reinalda (Ed.), *Routledge Handbook of International Organization*, 257–69. Abingdon-on-Thames, England: Routledge.

Egger, C., Picton, A., and Schopper, D. (2018). How do humanitarian IO conceptualise the world? An interdisciplinary approach to the analysis of concepts used in humanitarian action. Presented at the RUN Workshop 2018, Geneva, Switzerland. https://archive-ouverte.unige.ch/unige:135922

Faber, P. (2022). Frame-based terminology. In P. Faber, and M. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, Amsterdam, The Netherlands: John Benjamins, 353–76. https://doi.org/10.1075/tlrp.23.16fab

Hampton, J. A. (2020). Investigating differences in people's concept representations. In T. Marques and A. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*, 67–82. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198803331.003.0005

Hofmann, C., and Schneckener, U. (2011). Engaging non-state armed actors in state- and peace-building: Options and strategies. *International Review of the Red Cross* 93(883): 603–621. https://doi.org/10.1017/S1816383112000148

Horwood, G. V. (2017). Humanitarian assistance and disaster relief (HA/DR) articles and lexicon (Version V1) [dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/TGOPRU

Humanitarian Encyclopedia. (2021a). Corpus. *Humanitarian Encyclopedia*. https://humanitarianencyclopedia.org/corpus

———. (2021b). Methodology. *Humanitarian Encyclopedia*. https://humanitarianencyclopedia.org/methodology

———. (2021c). Overview. *Humanitarian Encyclopedia*. https://humanitarianencyclopedia.org/goals

IFRC. (2016). *World Disasters Report. Resilience: Savings Lives Today, Investing for Tomorrow*. Geneva, Switzerland: International Federation of Red Cross and Red Crescent Societies.

Isaacs, L. (2023a). Corpusama (0.1.1) [software]. https://github.com/engisalor/corpusama

———. (2023b). Humanitarian reports on ReliefWeb as a domain-specific corpus. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, 258–279, Brno, Czech Republic, June. Lexical Computing.

Isaacs, L., Odlum, A., and León-Araúz, P. (2024). Quartz: A template for quantitative corpus data visualization tools. *Languages*, 9(3): 81. https://doi.org/10.3390/languages9030081

Jolliffe, K. (2015). *Ethnic Armed Conflict and Territorial Administration in Myanmar*. The Asia Foundation. https://reliefweb.int/node/1726761

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431, Valencia, Spain, April. Association for Computational Linguistics (ACL).

Khan, A.K., and Kontinen T. (2022). Impediments to localization agenda: humanitarian space in the rohingya response in bangladesh. *Journal of International Humanitarian Action* 7(14). https://doi.org/10.1186/s41018-022-00122-1

Kovář, V., Močiariková, M., and Rychlý. P. (2016). Finding definitions in large corpora with Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 391–394, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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

León-Araúz, P., Faber, P., and Montero Martínez, S. (2012). Specialized language semantics. In P. Faber (Ed.), *A Cognitive Linguistics View of Terminology and Specialized Language*, Berlin: De Gruyter Mouton, 95–175.

León-Araúz, P., and San Martín, A. (2018). The EcoLexicon Semantic Sketch Grammar: From knowledge patterns to word sketches. In *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, 94–99, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Littell, P., Tian, T., Xu, R., Sheikh, Z., Mortensen, D., Levin, L., Tyers, F., Hayashi, H., Horwood, G., Sloto, S., Tagtow, E., Black, A., Yang, Y., Mitamura, T., and Hovy, E. (2018). The ARIEL-CMU situation frame detection pipeline for LoReHLT16: A model translation approach. *Machine Translation*, 32: 105–126. https://doi.org/10.1007/s10590-017-9205-3

Marshman, E. (2022). Knowledge patterns in corpora. In P. Faber, and M. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, Amsterdam, The Netherlands: John Benjamins, 291–310. https://doi.org/10.1075/tlrp.23.13mar

Miah, M. S. U., Sulaiman, J., Sarwar, T. B., Naseer, A., Ashraf, F., Zamli, K. Z., and Jose, R. (2022). Sentence boundary extraction from scientific literature of electric double layer capacitor domain: Tools and techniques. *Applied Sciences*, 12(3), 1352. https://doi.org/10.3390/app12031352

Michelfeit, J., Pomikálek, J., and Suchomel, V. (2014). Text tokenisation using unitok. In *RASLAN 2014 Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, 71–75, Karlova Studánka, Czech Republic, December. NLP Consulting.

Muis, A. O., Otani, N., Vyas, N., Xu, R., Yang, Y., Mitamura, T., and Hovy, E. (2018). Low-resource cross-lingual event type detection via distant supervision with minimal effort. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 70–82, New Mexico, USA, August. Association for Computational Linguistics (ACL).

Nemeskey, D. M., and Kornai, A. (2018). Emergency vocabulary. *Information Systems Frontiers*, 20(5): 909–923. https://doi.org/10.1007/s10796-018-9843-x

Padró, L., and Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2473–2479, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Papadopoulos, P., Travadi, R., Vaz, C., Malandrakis, N., Hermjakob, U., Pourdamghani, N., Pust, M., Zhang, B., Pan, X., Lu, D., Lin, Y., Glembek, O., Baskar, M. K., Karafiát, M., Burget, L., Hasegawa-Johnson, M., Ji, H., May, J., Knight, K., and Narayanan, S. S. (2017). Team ELISA system for DARPA LORELEI speech evaluation 2016. In *INTERSPEECH 2017*, 2053–2057, Stockholm, Sweden, August. ISCA. https://doi.org/10.21437/Interspeech.2017-180

OCHA. (2012). *OCHA Annual Report 2012*. Geneva, Switzerland: Office for the Coordination of Humanitarian Affairs.

OHCHR. (2021). *United Nations Human Rights Report 2020*. United Nations Human Rights Office of the High Commissioner. https://reliefweb.int/node/3973410

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108, online, April. Association for Computational Linguistics (ACL).

ReliefWeb. (2008). *Reliefweb glossary of humanitarian terms*. https://reliefweb.int/report/world/reliefweb-glossary-humanitarian-terms-enko

Rey Marcos, F. (2003). Reflexiones sobre el humanitarismo. *Papeles*, 82: 43–49.

Roepstorff, K. (2020). Localisation and shrinking civic space: Tying up the loose ends. *Centre for Humanitarian Action* (blog). https://www.chaberlin.org/en/publications/localisati

on-and-shrinking-civic-space-tying-up-the-loose-ends/
Rüttinger, L., Munayer, R., Ackern, P. van, and Titze, F. (2022). *The nature of conflict and peace. The links between environment, security and peace and their importance for the United Nations*. Gland, Swizerland/Berlin, Germany: WWF International/adelphi consult.
Rychlỳ, P. (2007). Manatee/Bonito—A modular corpus manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, 65–70, Brno, Czech Republic, December. Masaryk University.
San Martín, A., Trekker, C., and León-Araúz, P. (2020). Extraction of hyponymic relations in French with knowledge-pattern-based word sketches. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC-2020)*, 5953–5961, Marseille, France, May. European Language Resources Association (ELRA).
Sezgin, Z., and Dijkzeul D, (2015). *The new humanitarians in international practice: emerging actors and contested principles*. London, UK: Routledge. https://doi.org/10.4324/9781315737621
Shamoug, A., Cranefield, S., and 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., and Bach C. (2010). Definitional verb patterns for semantic relation extraction. In A. Auger, and C. Barrière (Eds.), *Probing Semantic Relations: Exploration and Identification in Specialized Texts*, 74–96. Amsterdam, The Netherlands: John Benjamins. https://doi.org/10.1075/bct.23.04sie
Tamagnone, N., Fekih, S., Contla, X., Orozco, N., and Rekabsaz, N. (2023). Leveraging domain knowledge for inclusive and bias-aware humanitarian response entry classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 6219–6227, Macau, China, December. IJCAI. https://doi.org/10.24963/ijcai.2023/690
Ward, V., Smith, S., House, A., and Hamer, S. (2012). Exploring knowledge exchange: A useful framework for practice and policy. *Social Science & Medicine*, 74(3): 297–304. https://doi.org/10.1016/j.socscimed.2011.09.021
WHS. (2016). Commitments to action. Istanbul, Turkey: World Humanitarian Summit.