Humanitarian Corpora for English, French and Spanish

Loryn Isaacs, Santiago Chambó & Pilar León Araúz

University of Granada, Dept. of Translation & Interpreting

{lisaacs;santiagochambo;pleon}@ugr.es

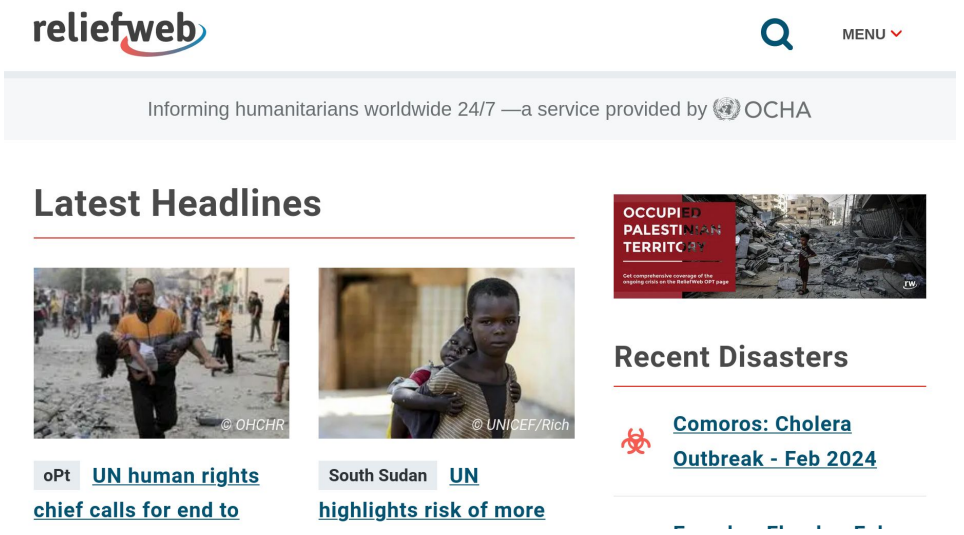
LREC-COLING 2024

What is ReliefWeb (RW)?

1. The largest database of humanitarian reports (1< million)

2. Humanitarians and others use it to
   - Coordinate humanitarian response
   - Study domain discourse & concepts

3. Requires adaptation for linguistic study
   - Full-text search
   - Quantitative analysis
   - Language disaggregation



https://reliefweb.int/

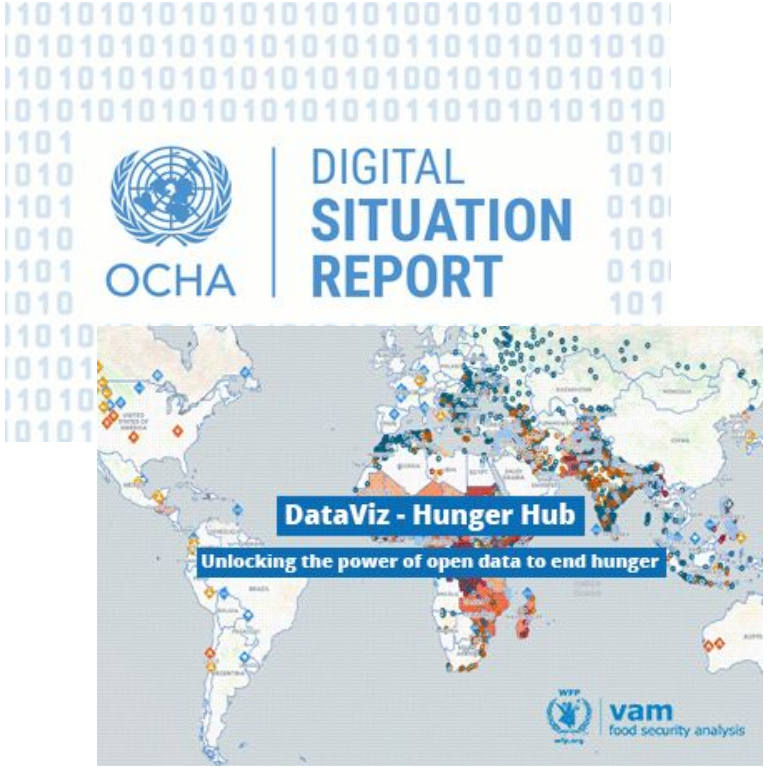
How is RW used?

Other projects

- Tracking famine discursively (Rubin, 2014)
- ReliefWeb corpora for ML (Horwood, 2017)
- Semantic embedding (Shamoug et al., 2023)
- HumBERT LLM (Tamagnone et al., 2023)
- 25+ projects on ReliefWeb Labs

Our focus

- Descriptive statistics of word frequencies
- Pattern-based knowledge extraction (León-Araúz & San Martín, 2018)
- Conceptual analysis (Chambó & León-Araúz, 2023)



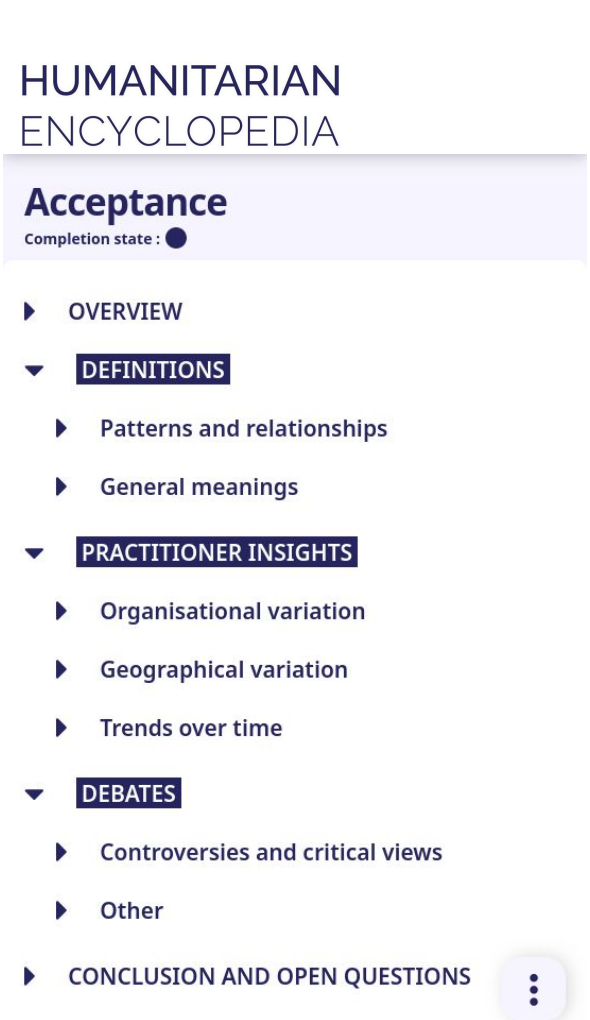
https://labs.reliefweb.int/

Our work

Humanitarian Encyclopedia & University of Granada LexiCon research group collaboration

- Conduct corpus-based analyses on concepts (humanitarianism, gender-based violence, independence, …)
- Explore ill-defined or debated concepts
- Contribute to the shared understand of domain's key concepts
- Expand data resources & management
- Produce open-source analysis tools

https://humanitarianencyclopedia.org

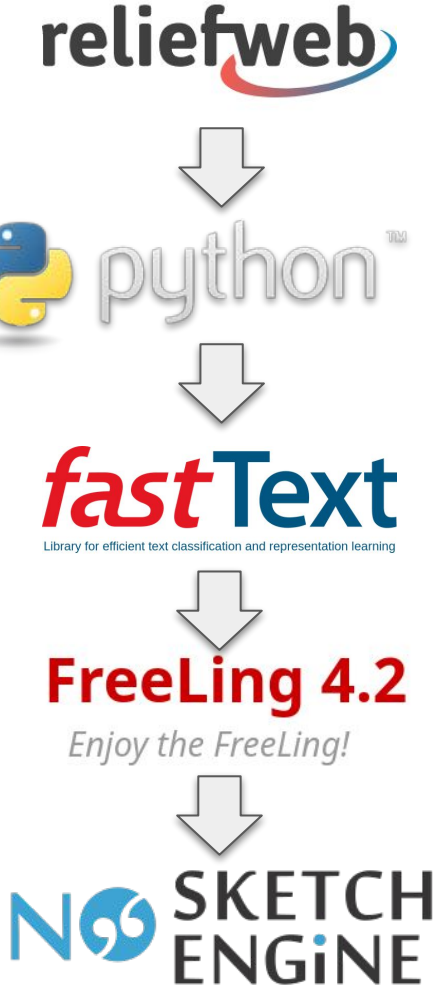


Generating ReliefWeb corpora

(Summary)

1. Download ReliefWeb data (HTML & PDF) via API
2. Clean & standardize texts
3. Identify language(s) with fastText (Abadji et al., 2022; Joulin et al., 2017)
4. Disaggregate & reduce noise
5. Process with FreeLing NLP (tokenization, lemmatization, POS) (Padró & Stanilovsky, 2012)
6. Load into NoSketch Engine corpus software (Kilgarriff et al., 2014; Rychlỳ, 2007)

Expanded from Isaacs (2023)
https://github.com/engisalor/corpusama



Language identification results

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

Corpus composition

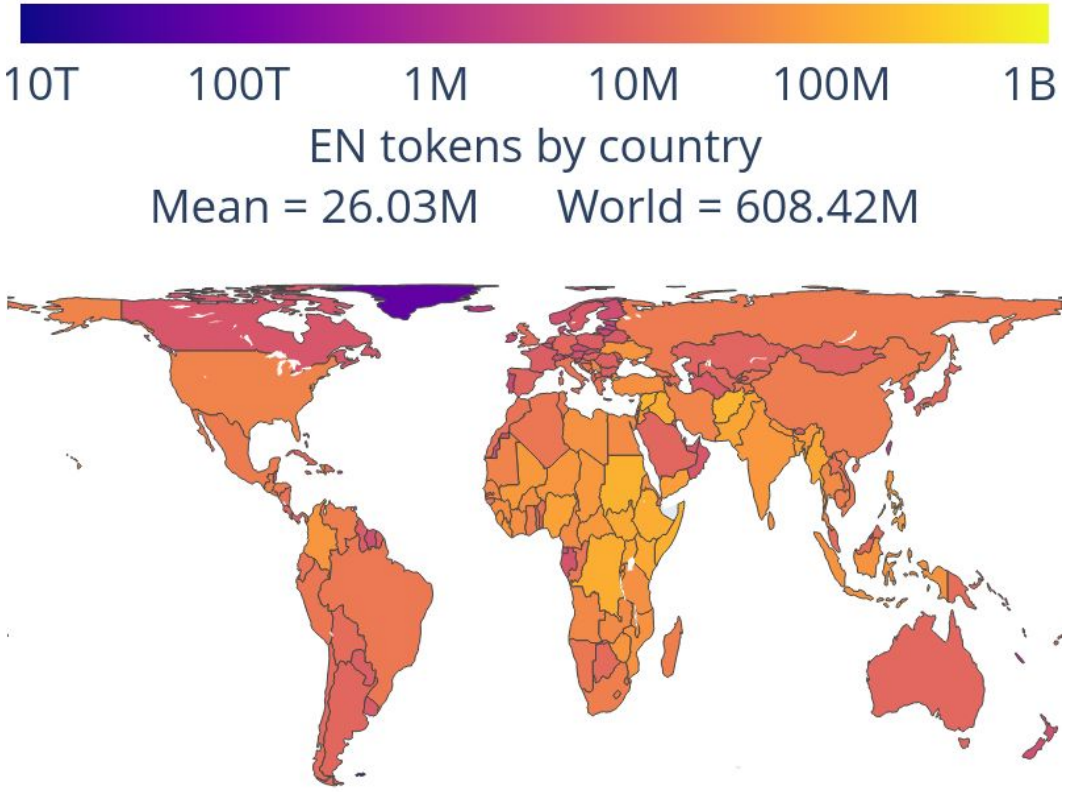
Corpus	Docs	Types M	Tokens M	PDF tokens
EN	858,657	1,608	1,983	78.25%
FR	104,602	196	235	74.26%
ES	76,919	118	142	77.80%
Total	*1,040,178*	*1,922*	*2,360*	*76.77%**

*Mean

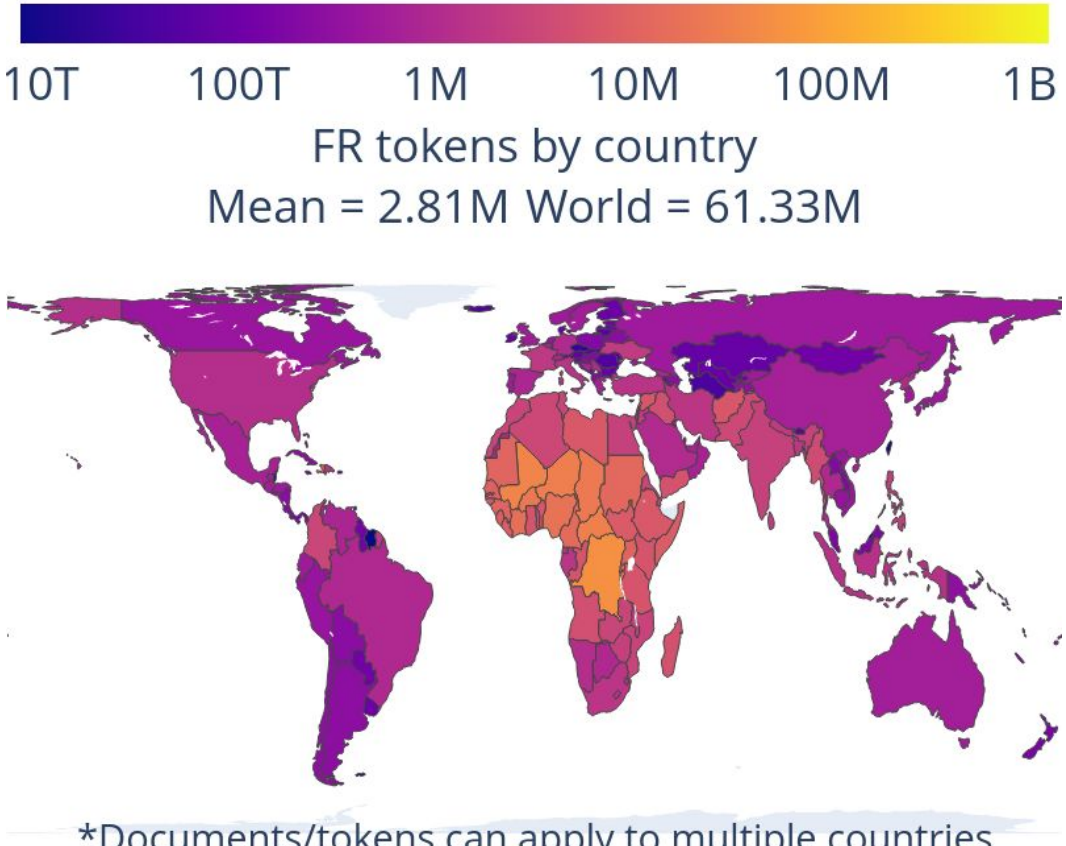
Key text types

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

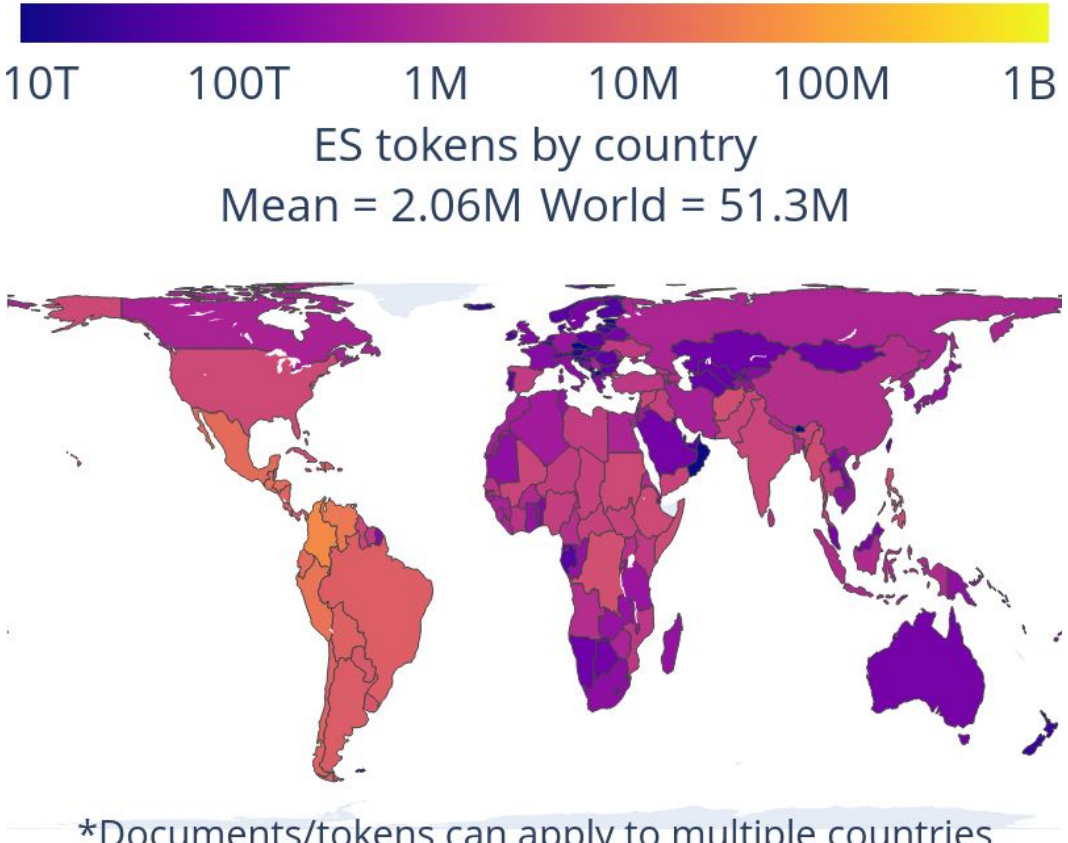
Geographic distribution: English



Geographic distribution: French



Geographic distribution: Spanish



Probing the corpus: exploratory analysis

Concept of ARMED ACTOR

Pattern-based knowledge extraction (León-Araúz & San Martín, 2018)

- Definitions
- KRCs (Marshman, 2022) containing generic-specific information
  - Hypernyms
  - Hyponyms
  - MWT hyponyms

Concept	HE Corpus	ReliefWeb EN
Armed actors	**299**	**15414**
Responsibility-to-protect	317	16058
Do no harm	365	12571
Sanitation	25059	566296
Advocacy	27266	198652

Definition extraction for ARMED ACTOR

Extraction with definitional verbal and paralinguistic patterns (Sierra et al., 2010; Dorantes et al., 2017)

No definitions in ReliefWeb ES
8 definitions in ReliefWeb EN

- 7 with ‘non-state armed actor’ and ‘ethnic armed actor’ as difinienda
- ARMED ACTOR as collective entities denoting organized human groups
- Definitional effort focuses on ‘non-state armed actor’

Hypernyms and hyponyms of ARMED ACTOR (1)

Targets:

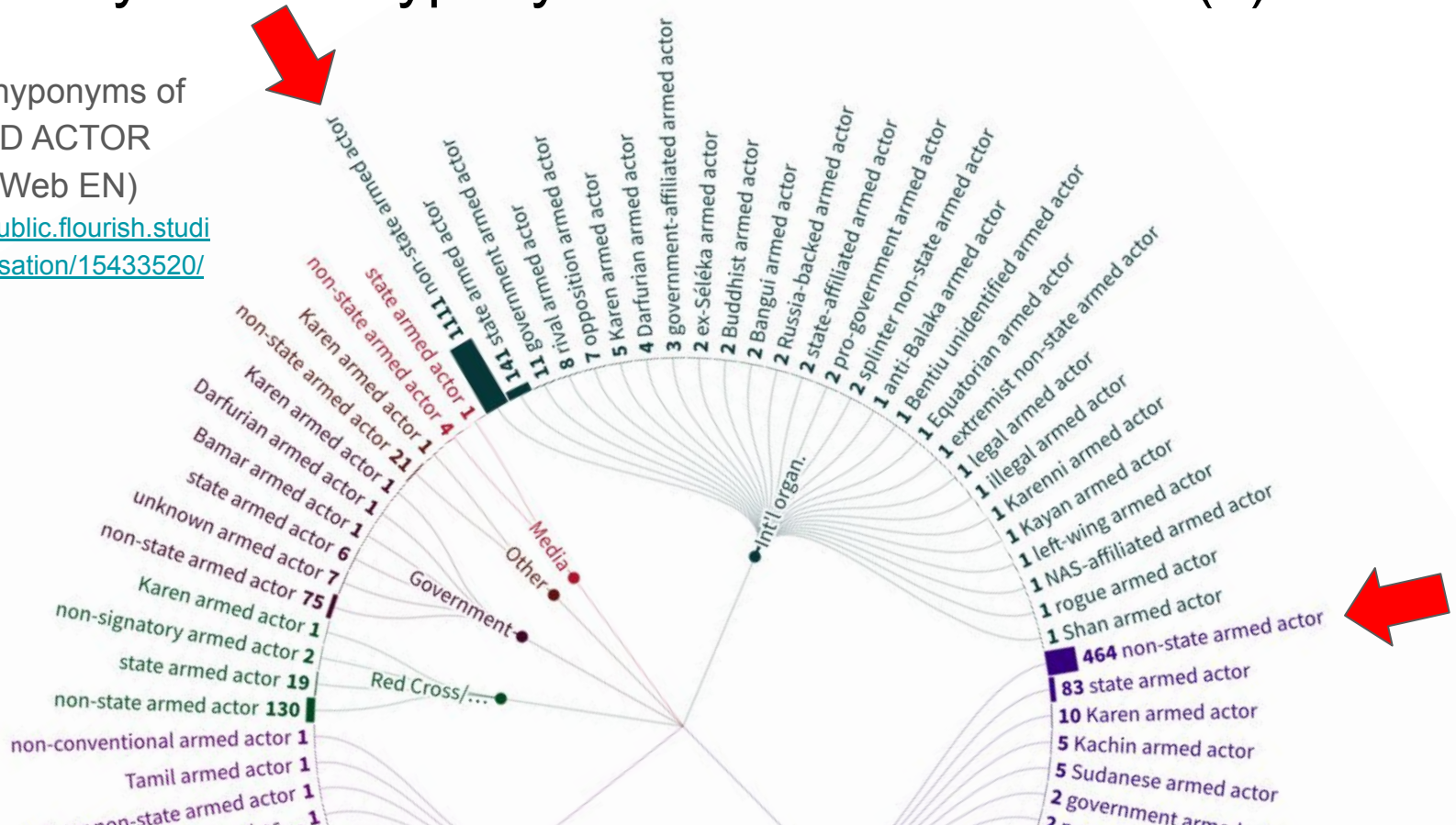
- Generic-specific KRCs
  - guerilla is a type of armed actor
  - armed actors include drug-traffickers

- MWT hyponyms
  Important because semantic relations can be inferred from internal structure (Cabezas-García & León-Araúz, 2018)
  - government-affiliated armed actor
  - cross-border armed actor

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)

Hypernyms and hyponyms of ARMED ACTOR (2)

MWT hyponyms of ARMED ACTOR (ReliefWeb EN)
https://public.flourish.studio/visualisation/15433520/



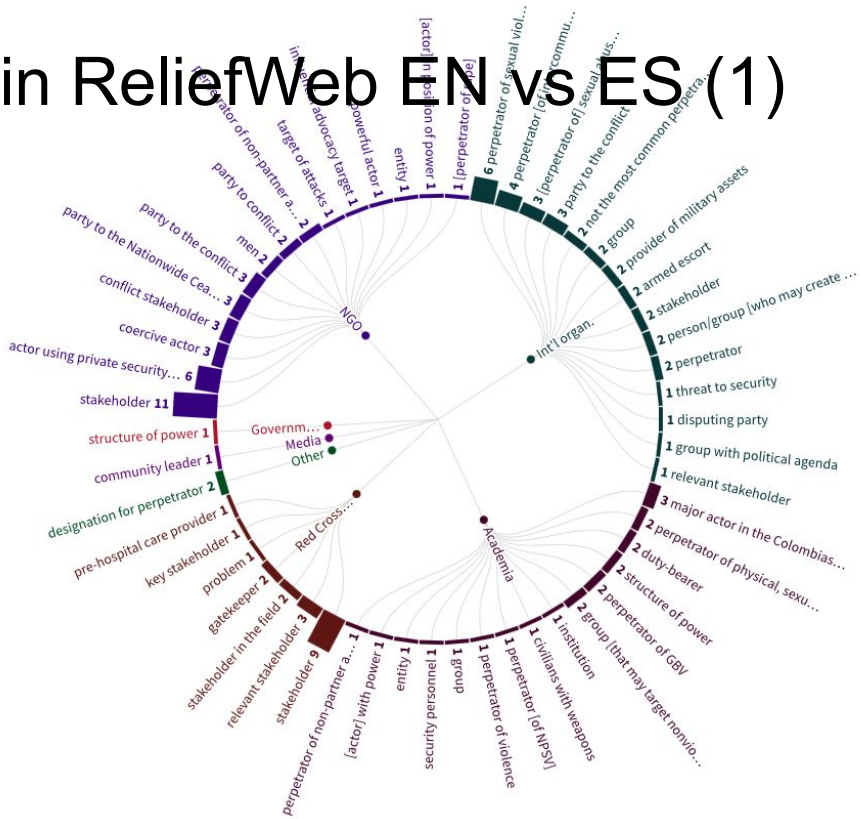
Comparing ARMED ACTORS in ReliefWeb EN vs ES (1)

Hypernyms:

- Not as abundant in ES as in EN

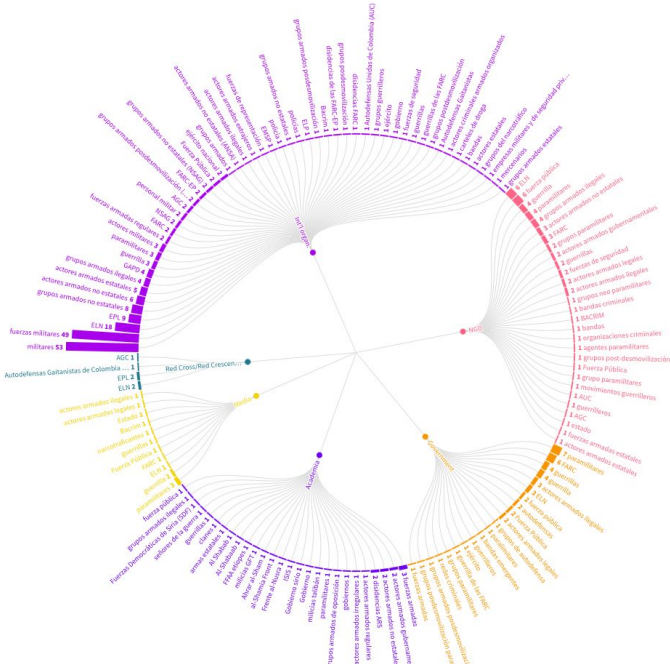
Hyponyms:

- Colombian named entities
  - Fuerzas Armadas Revolucionarias de Colombia (FARC)
  - Ejército de Liberación Nacional (ELN)
  - Autodefensas Unidas de Colombia (AUC)
  - Autodefensas Gaitanistas de Colombia (AGC)



Hypernyms of ARMED ACTOR (ReliefWeb EN)
https://public.flourish.studio/visualisation/15433955/

Comparing ARMED ACTORS in ReliefWeb EN vs ES (2)



Hyponyms of ARMED ACTOR (ReliefWeb ES)
https://public.flourish.studio/visualisation/15438038/

Hyponyms:

- International organizations:
  - EN: ‘military’ and ‘military forces’
  - ES: ‘fuerzas militares’ and ‘militares’
- Academia:
  - EN: rebel groups, ‘Israeli armed forces’, ‘police’ and ‘government’
  - ES: Colombian armed forces and ‘actores armados gubernamentales’
- Government and NGO:
  - EN: ‘guerrilla’ and ‘paramilitary group’
  - ES: ‘grupos paramilitaries’ and Colombian named entities

Comparing ARMED ACTORS in ReliefWeb EN vs ES (3)

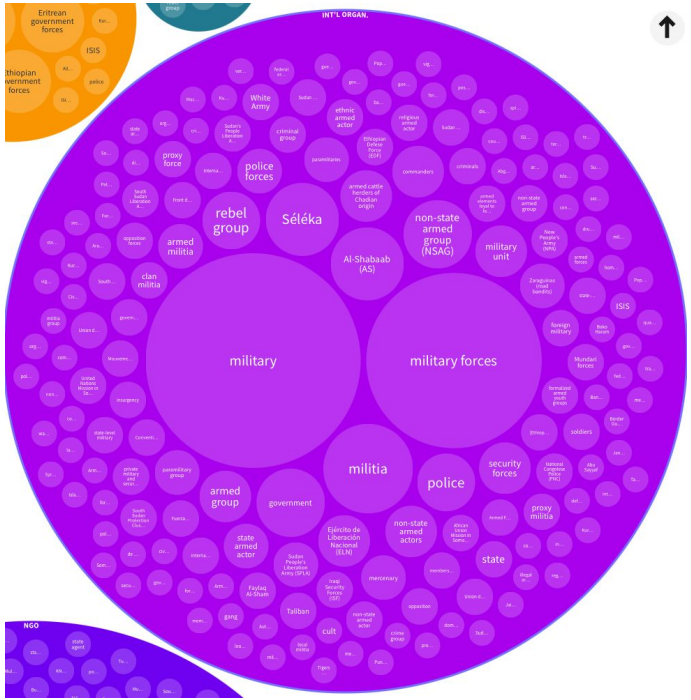
MWT hyponyms:

- ‘actor armado ilegal’ [illegal armed actor]
- Red Cross and Academia:
  - ES: ‘actor armado no estatal’ [non-state actor]

ARMED ACTOR in ReliefWeb ES

- Colombia-centric understanding
- ‘actor armado ilegal’ vs ‘non-state actor’

“In Colombia, the many non-state armed groups are mostly referred to as illegal armed actors.” (Rüttinger et al., 2022: 24)



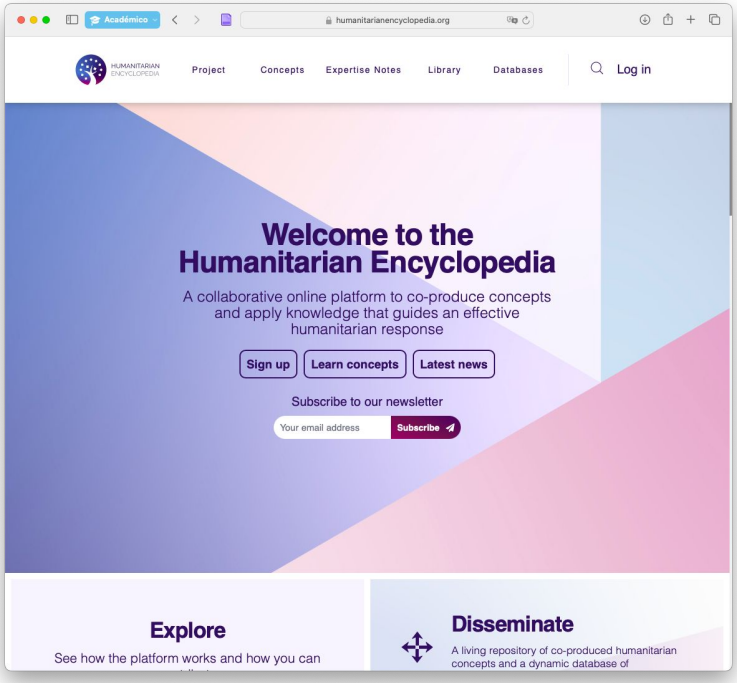
Hyponyms of ARMED ACTOR
(international organizations in ReliefWeb EN)
https://public.flourish.studio/visualisation/15437412/

Conclusions

1. New larger and up-to-date corpora for the Humanitarian Encyclopedia

2. Exploratory analysis on ARMED ACTOR

3. Multilingual corpora for studies on interlinguistic conceptual variation and languages other than English



https://humanitarianencyclopedia.org

Acknowledgements

Funding

- PROYEXCEL_00369 (VariTermiHum)
  Regional Government of Andalusia

Acknowledgements

- Humanitarian Encyclopedia

UNIVERSIDAD DE GRANADA

HUMANITARIAN ENCYCLOPEDIA

References

Abadji, J., Ortiz Suarez, P., Romary, L., & 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.

Cabezas-García, M., & León-Araúz, P. (2018). Towards the Inference of Semantic Relations in Complex Nominals: A Pilot Study. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1399

Chambó, S., & León-Araúz, P. (2023). Operationalising and representing conceptual variation for a corpus-driven encyclopaedia. *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2023 Conference*, 587–612, Brno, Czech Republic.

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

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

Isaacs, L. (2023). Humanitarian reports on ReliefWeb as a domain-specific corpus. In M. Medveď, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, & M. Jakubíček (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference*, 248–269, Brno, Czech Republic.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431, Valencia, Spain.

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). 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 *Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets"*, 94–99, Miyazaki, Japan.

References

Marshman, E. (2022). Knowledge patterns in corpora. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* (Vol. 23, pp. 291–310). John Benjamins. https://doi.org/10.1075/tlrp.23.13mar

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 2473–2479, Istanbul, Turkey.

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

Rüttinger, L., Munayer, R., Ackern, P. van, & Titze, F. (2022). *The nature of conflict and peace. The links between environment, security and peace and their importance for the United Nations*. WWF International/adelphi consult GmbH. https://climate-diplomacy.org/sites/default/files/2022-05/WWF-adelphi_The%20Nature%20of%20Conflict%20and%20Peace_mid%20res_0.pdf

Rychlý, 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.

Sierra, G., Alarcón, R., Aguilar, C., & Bach, C. (2010). Definitional verbal patterns for semantic relation extraction. In *Probing Semantic Relations*. John Benjamins. https://www.jbe-platform.com/content/books/9789027287922-bct.23.04sie

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

Tamagnone, N., Fekih, S., Contla, X., Orozco, N., & 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*, 6219–6227, Macau, China. https://doi.org/10.24963/ijcai.2023/690