# Humanitarian reports on ReliefWeb as a domain-specific corpus

eLex 2023, Wed June 28

Loryn Isaacs - University of Granada
lisaacs@ugr.es

# What is ReliefWeb?

- 25+ year-old service
- 10.8 million users
- 1 million reports

"the principal information system for prevention, preparedness, and rapid response for the humanitarian community"

(Naidoo, 2007; OCHA, 2022; Ruso, 1996, p.18; Wackernagel & Footner, 2021)

https://www.unocha.org/

https://reliefweb.int/

# ReliefWeb’s value for linguists

- ReliefWeb Labs tracking tools
- Full API access
- Discursively track famine
- Knowledge extraction via semantic embedding

https://labs.reliefweb.int/

(Rubin, 2014; Shamoug, Cranefield & Dick, 2023)

# Shared goals

- Improve humanitarian response by leveraging linguistic data

- Address challenges synthesizing & transmitting domain knowledge

# Objectives

- Develop corpus from API data
- Analyze database's composition
- Gauge suitability & limits
- Facilitate research in humanitarian knowledge extraction

# Immediate applications

- Humanitarian Encyclopedia (HE)
  - 129 humanitarian concepts
  - Foster expert, linguist & community dialogue
  - Study multidimensionality

https://humanitarianencyclopedia.org/

(Chambó & León-Araúz, 2021; León-Araúz, 2017)

# Background

- Effect-size keyness analysis
  - RW & HE corpora
- Knowledge-rich contexts (KRCs)
- Frame-Based Terminology

(Condamines, 2022; Faber, 2022; Gabrielatos, 2018; Kilgarriff, 2012; Marshman, 2022;  Meyer, 2001; Sierra et al., 2008)

- Corpus query language
- Sketch Engine API wrapper: *Sketch Grammar Explorer*

(Isaacs, 2022; Jakubíček et al., 2010; León-Araúz & San Martín, 2018; San Martín et al., 2020)

github.com/engisalor/sketch-grammar-explorer

# Analysis

## Main concern

Are ReliefWeb reports (short HTML docs, incl. summaries) a good source of data for studying humanitarian concepts?

## Procedure

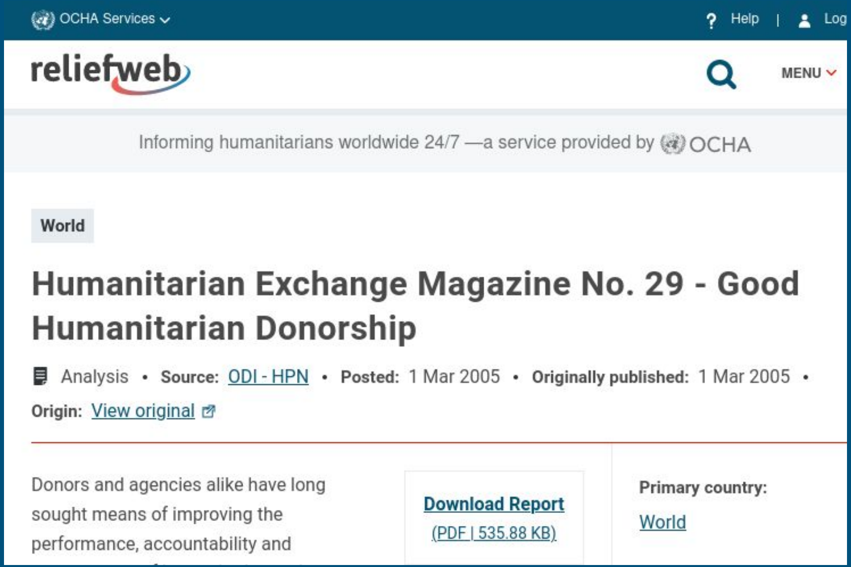
Compare ReliefWeb (RW) corpus w/ Humanitarian Encyclopedia (HE)

1. Concept keyness
2. Concept frequencies (annual)
3. KRCs (hypernymic & definitional)

# Corpus creation

- 662,473 reports (67%)
- English HTML content (PDFs excluded)
- Python code available on GitHub: *Corpusama*

https://github.com/engisalor/corpusama



Informing humanitarians worldwide 24/7 —a service provided by OCHA

World

Humanitarian Exchange Magazine No. 29 - Good Humanitarian Donorship

Analysis • Source: ODI - HPN • Posted: 1 Mar 2005 • Originally published: 1 Mar 2005 • Origin: View original

Donors and agencies alike have long sought means of improving the performance, accountability and

Download Report (PDF | 535.88 KB)

Primary country: World

https://reliefweb.int/node/23456

# Corpus creation

- Stanza NLP pipeline
- Universal Dependencies EWT treebank
- NoSketch Engine Docker

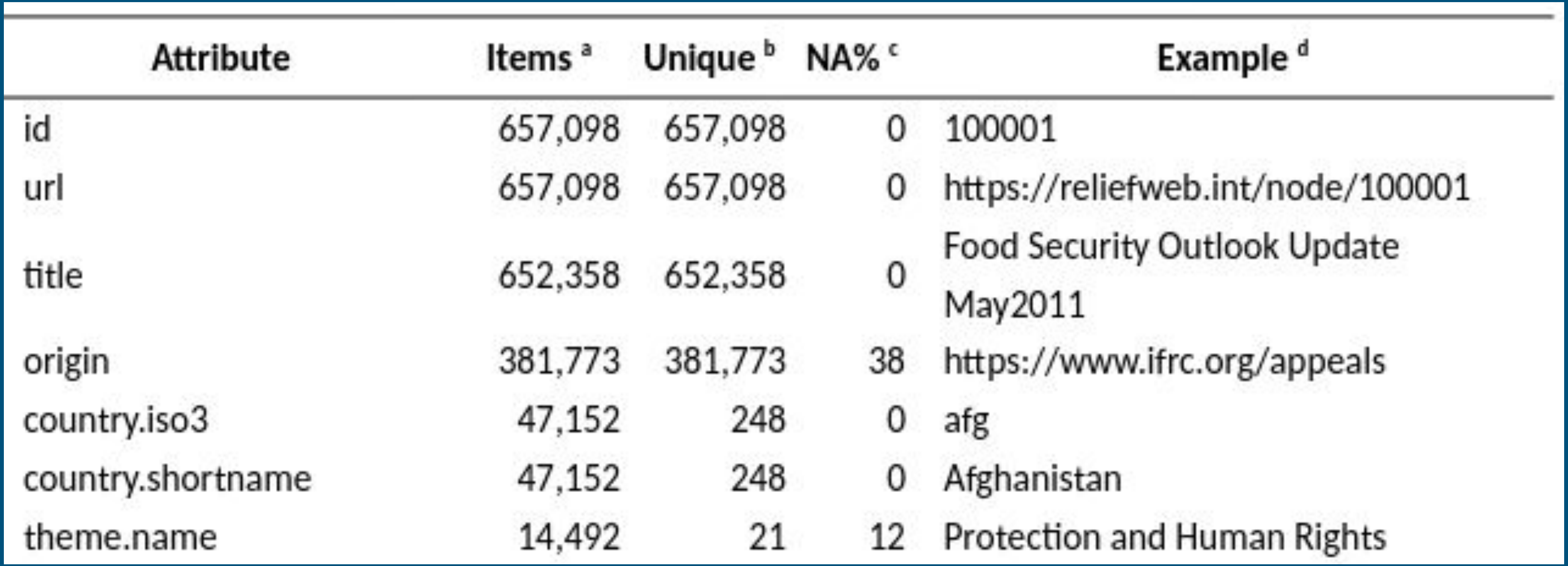
(ELTE-DH, 2023; Kilgarriff et al., 2014; Qi et al., 2020; Rychlỳ, 2007; Silveira et al., 2014)

# Corpus composition

- 657,098 documents
- 431,170,905 tokens
- 366,049,459 words
- 16,809,660 sentences
- 557 words/doc avg.

- Selection of API fields
- Preserve data structure
  - Multivalue fields (e.g., multiple affected countries)

# Corpus composition



| 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 |

# Corpus composition

- 248 countries (incl. “World”)
- Focus on African & Eastern Mediterranean WHO regions
- 2,708 organizations, esp. UN-related

- Natural disasters
  - flood, epidemic
- Themes
  - human rights, health, food & nutrition
- Genres
  - news, press releases & situation reports

# Keyness analysis

- 129 Humanitarian Encyclopedia concepts
- Compared shared years for RW & HE corpora
- K>1 count=10
- K<0.5 count=87
- K<0.25 count=38

## Keyness for shared years

| | | |
|---|---|---|
| min | <= | 0.006 |
| Q1 | = | 0.232 |
| Q2 | = | 0.358 |
| Q3 | = | 0.552 |
| max | = | 2.199 |

# Diachronic trends

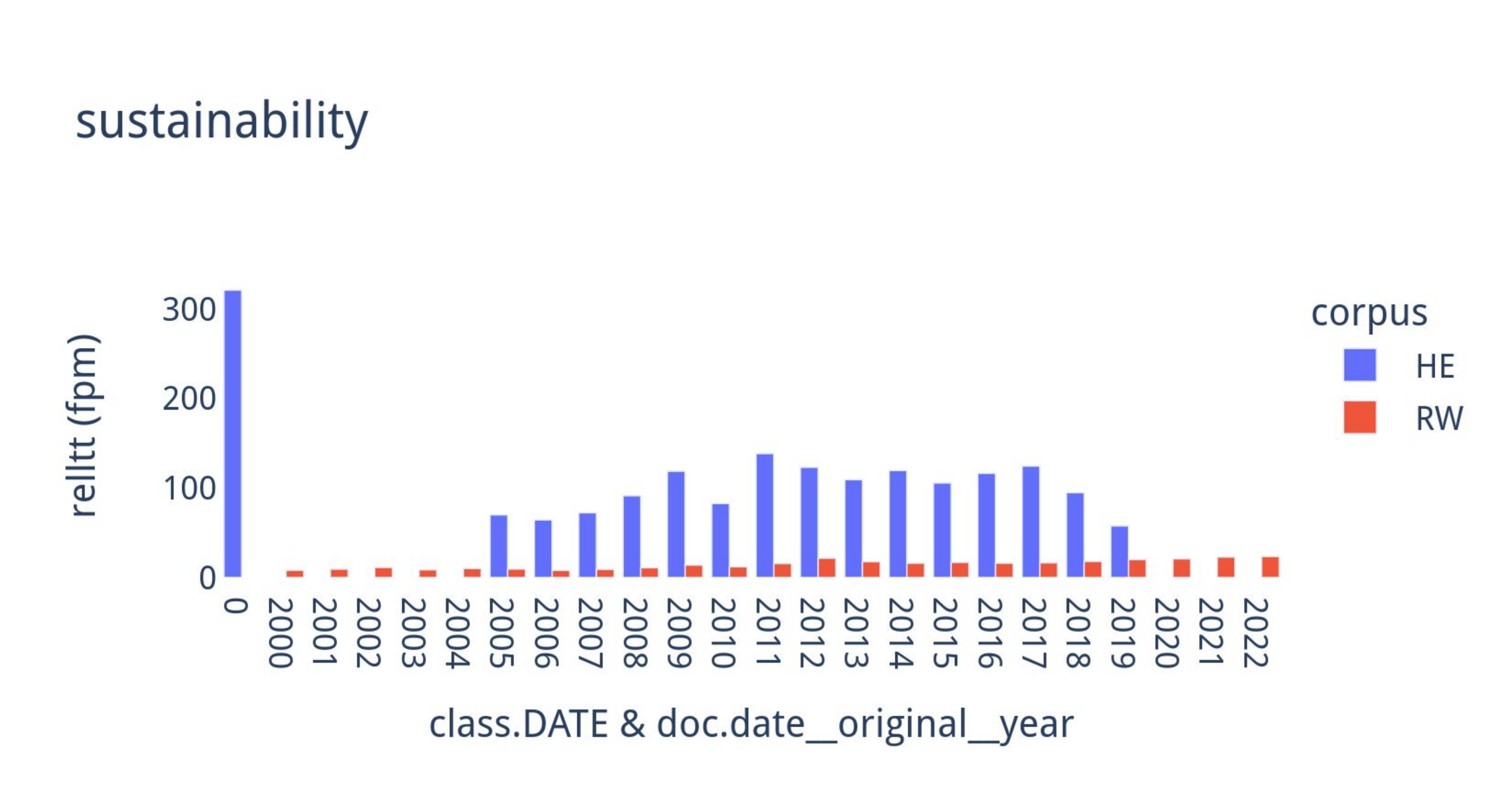
## Selected concepts

- HUMANITARIAN REFORM
- SUSTAINABILITY
- RESILIENCE
- GENDER-BASED VIOLENCE
- SETTLEMENT
- SOVEREIGNTY

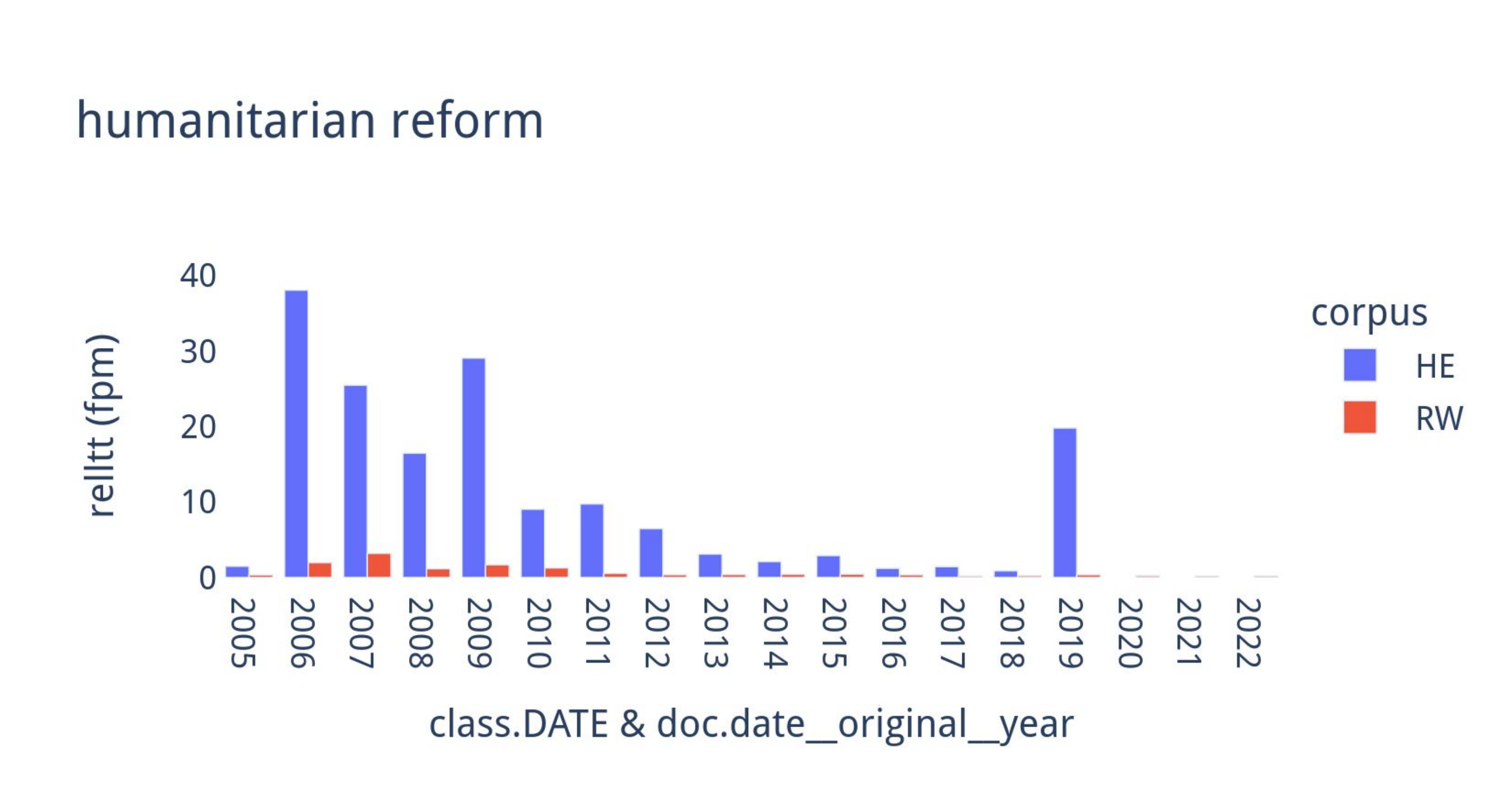
# Diachronic trends

- Corpora cover similar time period (~2000-2020)
  - Additional RW years inflate whole-corpus keyness
- Overlap between corpora?
- HE tagging irregularities
- Often flat or upward trends

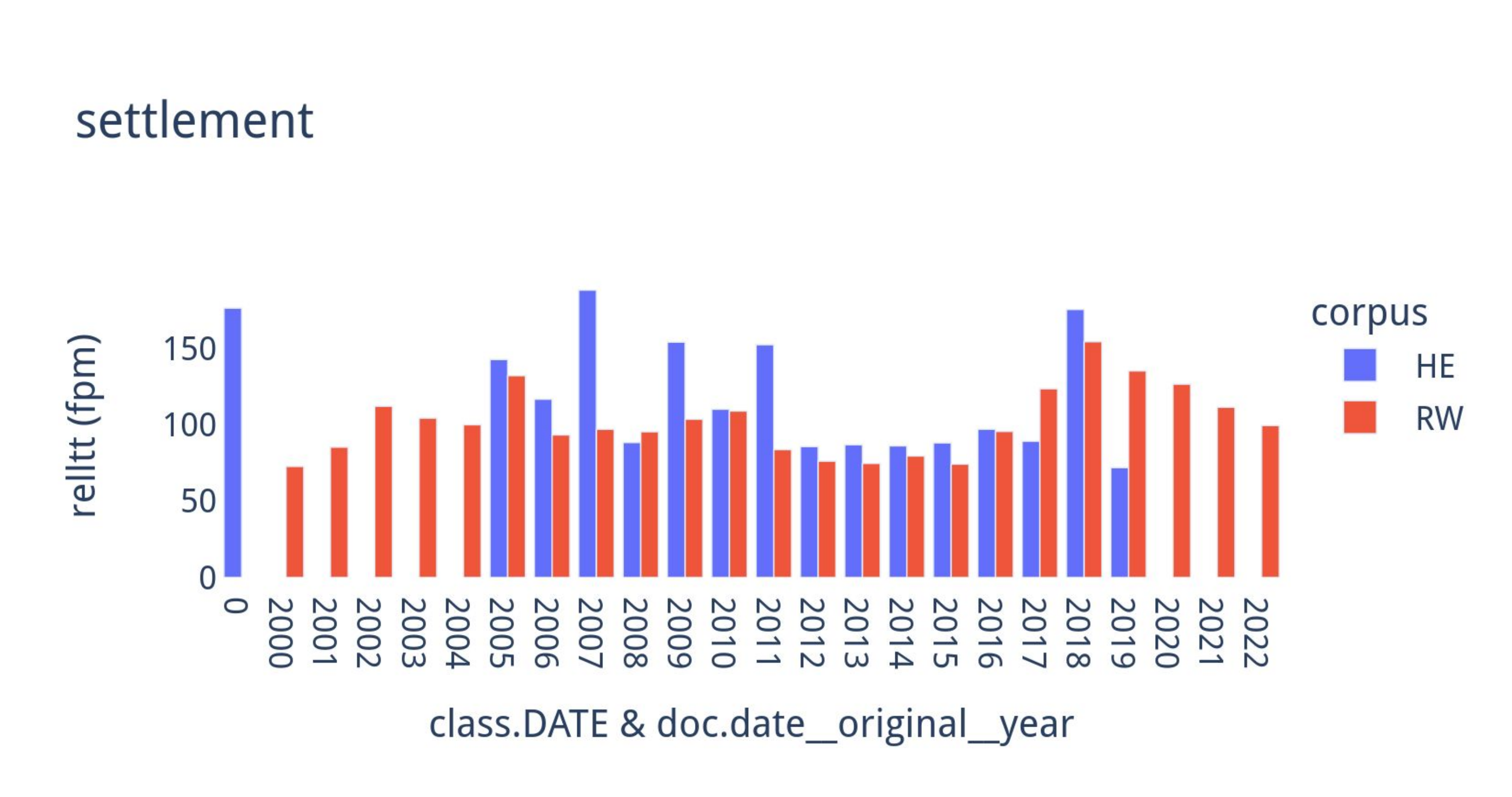
# Concepts with K<Q1



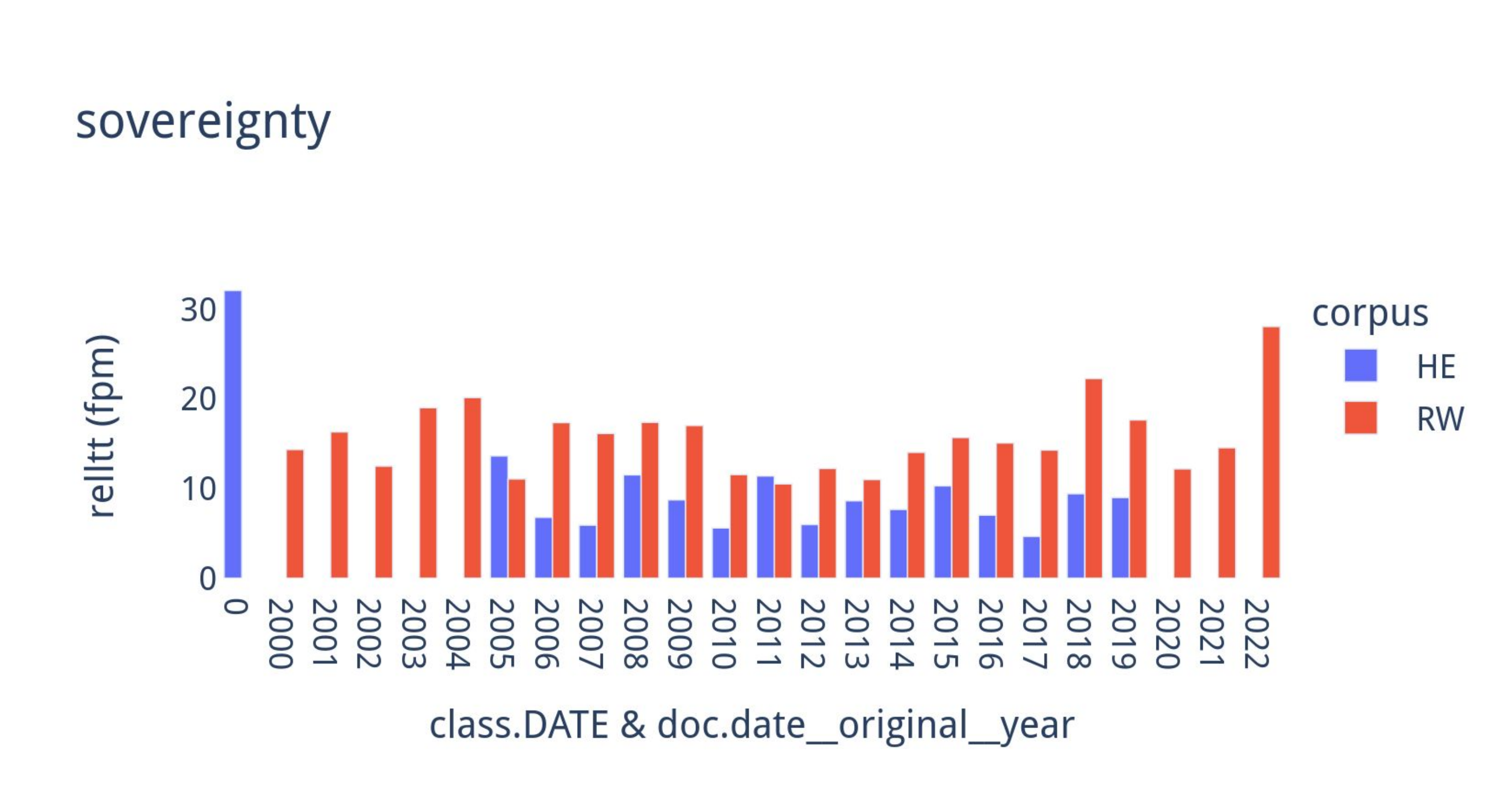
# Concepts with K<Q1



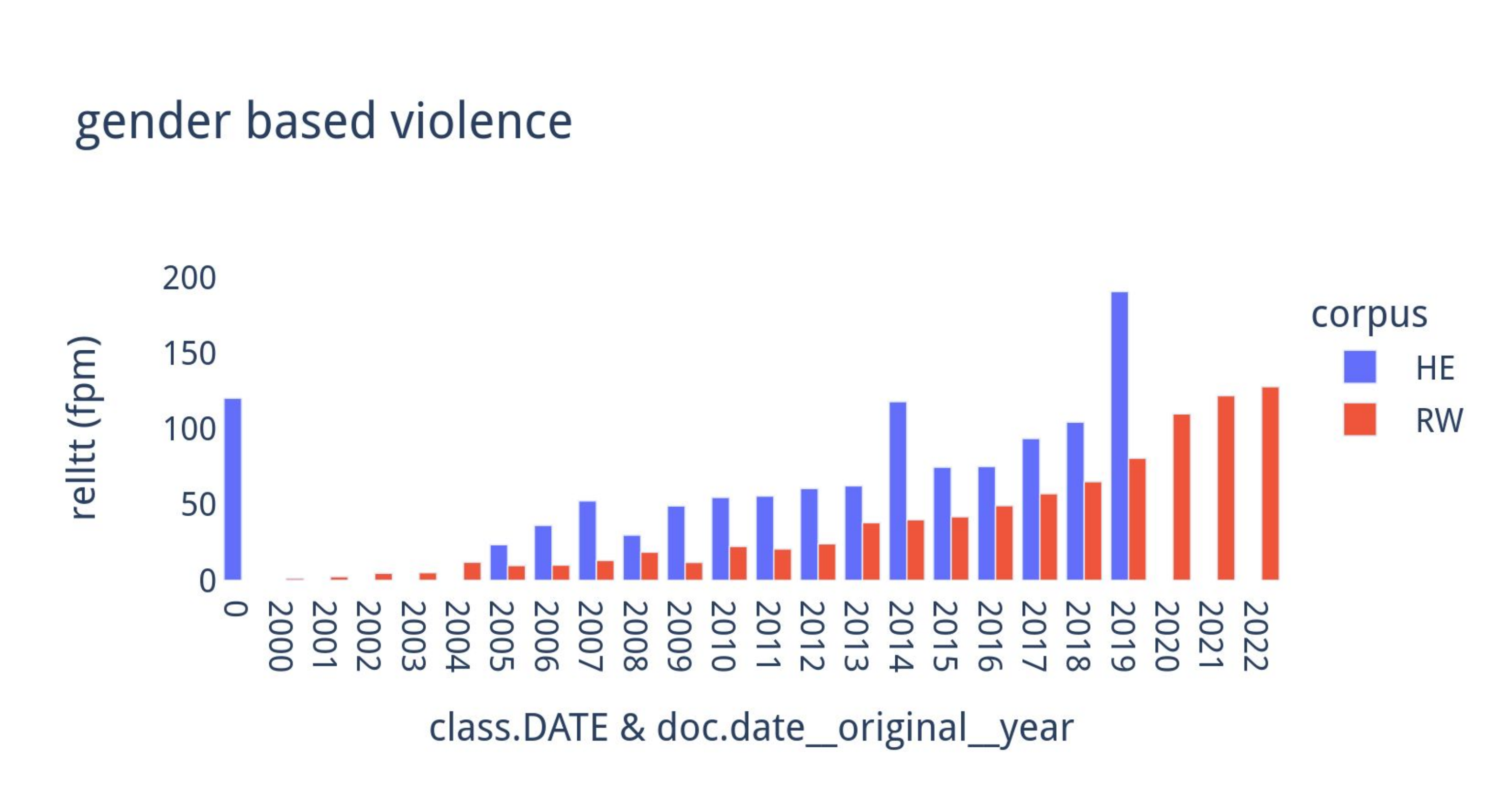
# Concepts with K>Q3



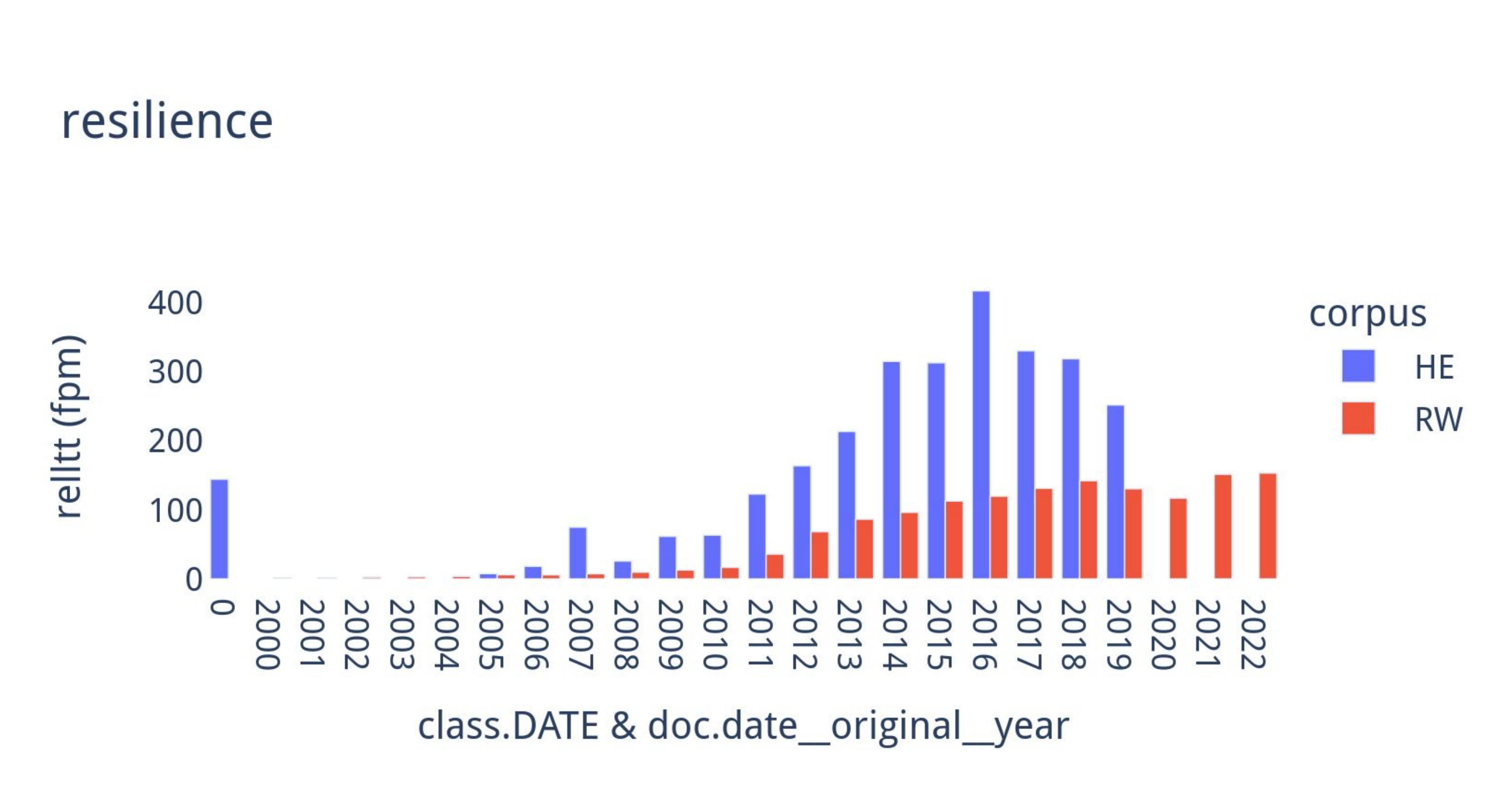
# Concepts with K>Q3



# Concepts with upward trends



# Concepts with upward trends



# KRC analysis

## Example KRC

**“Resilience is also a contested term in the literature”** (RW corpus)

| **Hyponym** | **Relation** | **Hypernym** |
|---|---|---|
| **resilience** | **IS_A** | **term** |

## Procedure

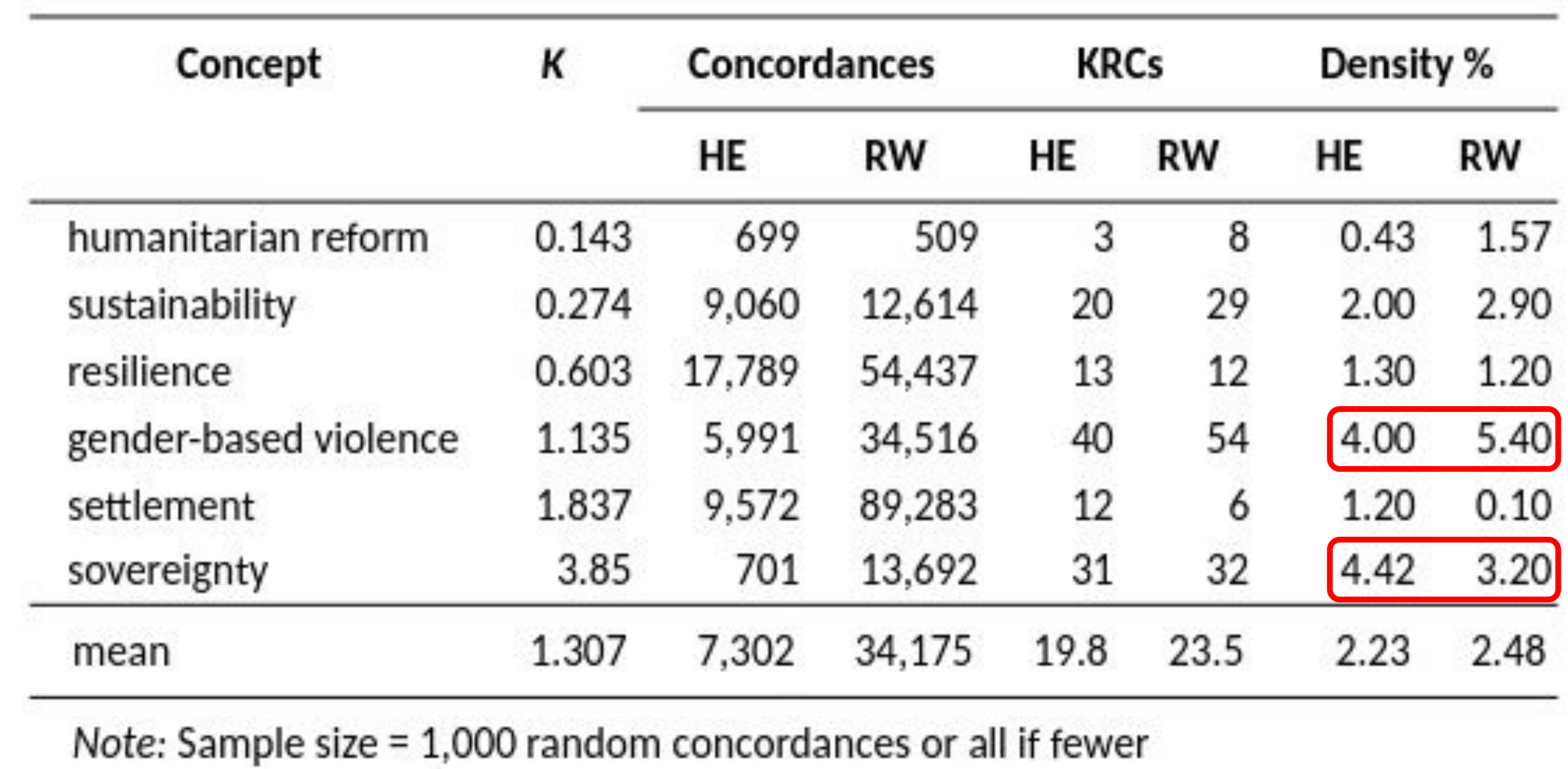
For sample concepts, compare

1. KRC density
2. Hypernyms (shared/unique)
3. Definitional contexts

# KRC density

- Samples of <=1,000 concordances
- Similar avg. densities, RW slightly higher
- RW = 2.48%
- HE = 2.23%

# KRC density



| 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 |

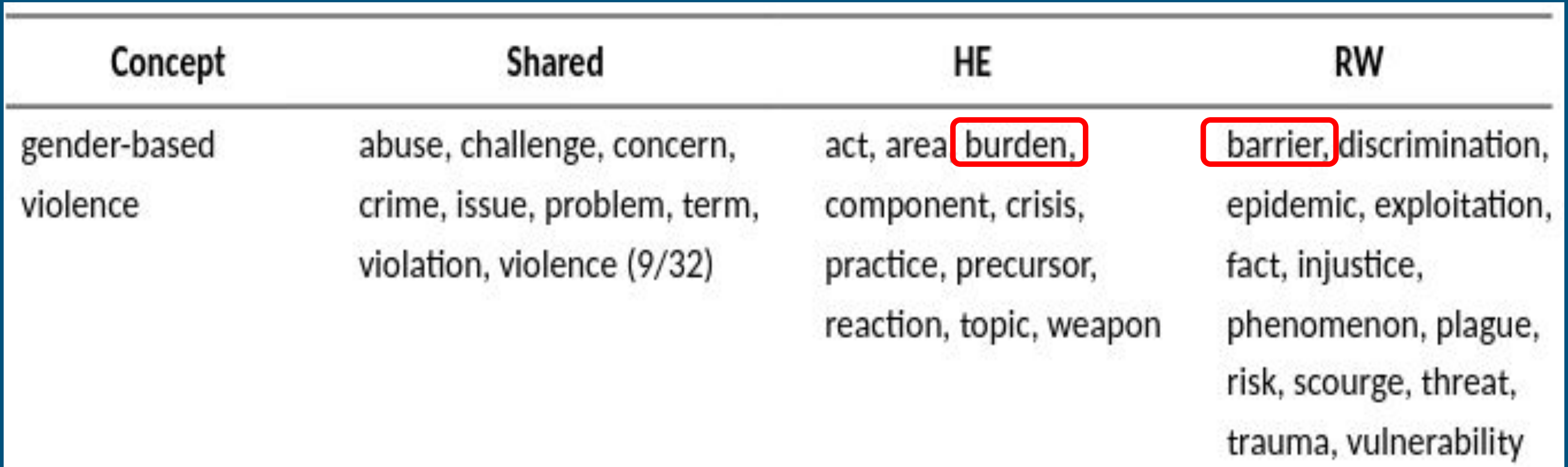
*Note*: Sample size = 1,000 random concordances or all if fewer

# Shared hypernyms

- 260 contexts extracted
- 104 hypernyms
- 24% shared across corpora

- HE = 34 unique
- RW = 45 unique

# Shared hypernyms



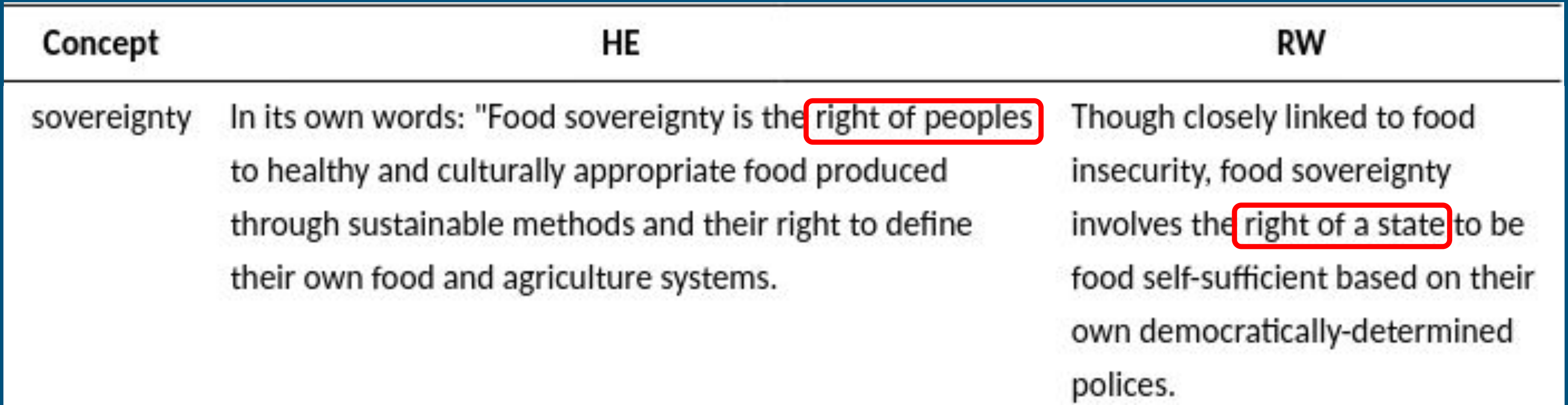
| Concept | Shared | HE | RW |
|---|---|---|---|
| gender-based violence | abuse, challenge, concern, crime, issue, problem, term, violation, violence (9/32) | act, area, burden, component, crisis, practice, precursor, reaction, topic, weapon | barrier, discrimination, epidemic, exploitation, fact, injustice, phenomenon, plague, risk, scourge, threat, trauma, vulnerability |

# Definitional contexts

- Only 3 concepts w/ definitions

| | |
|---|---|
| RESILIENCE | More contextualized in RW |
| GENDER-BASED VIOLENCE | Similar |
| SOVEREIGNTY | Similar; FOOD SOVEREIGNTY |

# Definitional contexts



| Concept | HE | RW |
|---|---|---|
| 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. |

# Discussion

- Low concept frequencies in RW HTML texts

↓

- Similar diachronic & KRC results

- More EVENTS afflicting populations
  - EPIDEMIC, FAMINE, …

↓

- Fewer humanitarian PROCESSES
  - ADVOCACY, SUSTAINABILITY, …

# Next steps

- Full concept analyses
- Study variation in more text types
- Underway:
  - Complete corpora (HTML & PDF)
  - Multiple languages (EN, FR, ES, ?)
  - HE Concept Tracker

    https://humanitarianencyclopedia.org/analysis

# Acknowledgements

UNIVERSIDAD
DE GRANADA

LexiCon
research group

## Funding

- Humanitarian Encyclopedia
  (Geneva Centre of Humanitarian Studies)
- PROYEXCEL_00369 (VariTermiHum)
  Regional Government of Andalusia

## Special thanks

- Open-source software, including
  - NoSketch Engine
  - Stanza NLP
  - ELTE-DH

# 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

# References (cont.)

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.

# References (cont.)

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

Rychlỳ, 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

# References (cont.)

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.

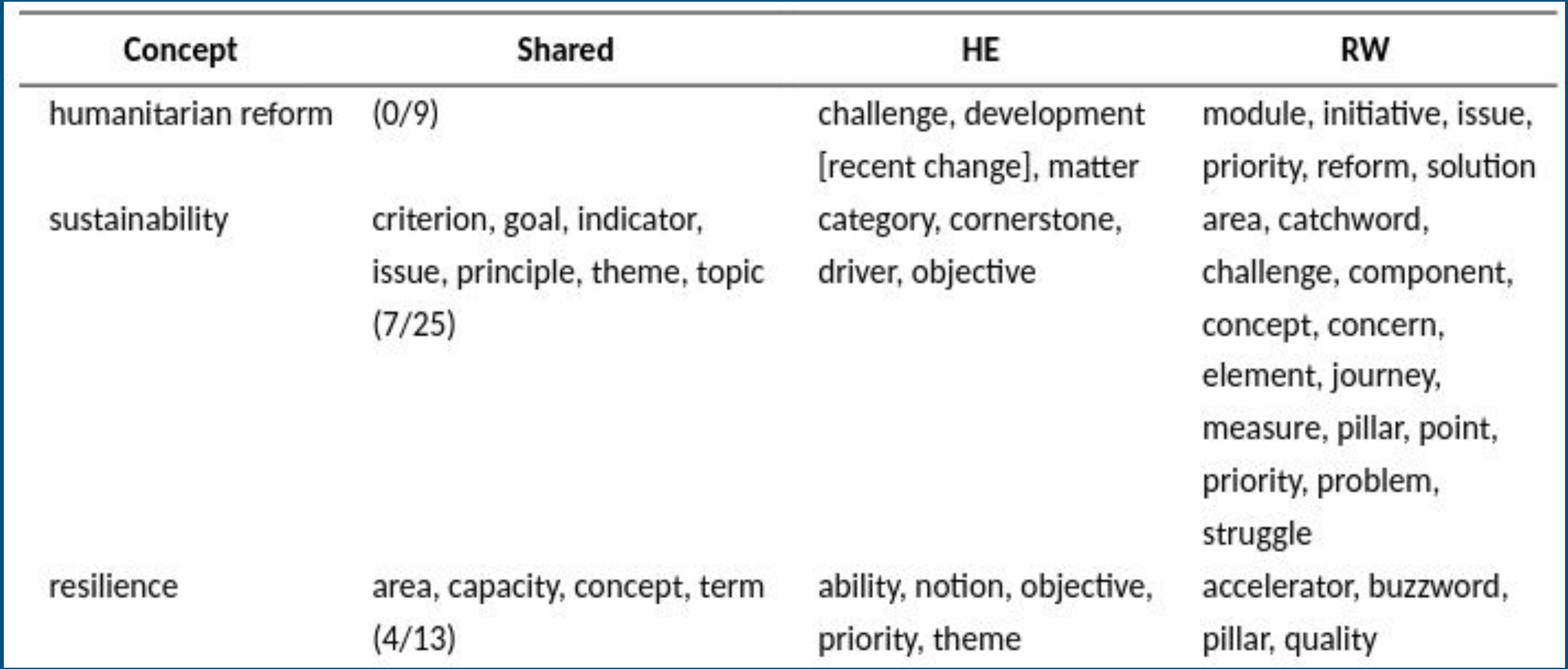
United Nations Office for the Coordination of Humanitarian Affairs. (2022). ReliefWeb analytics: 2022 mid-year highlights. https://reliefweb.int/report/world/reliefweb-highlights-mid-year-2022

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

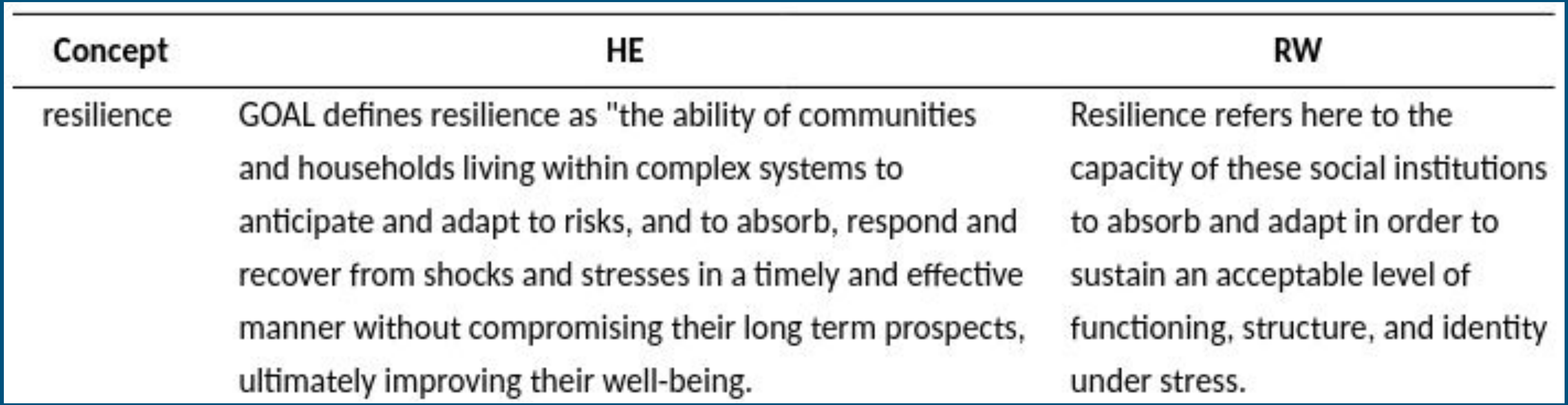
# Appendix

# Shared hypernyms



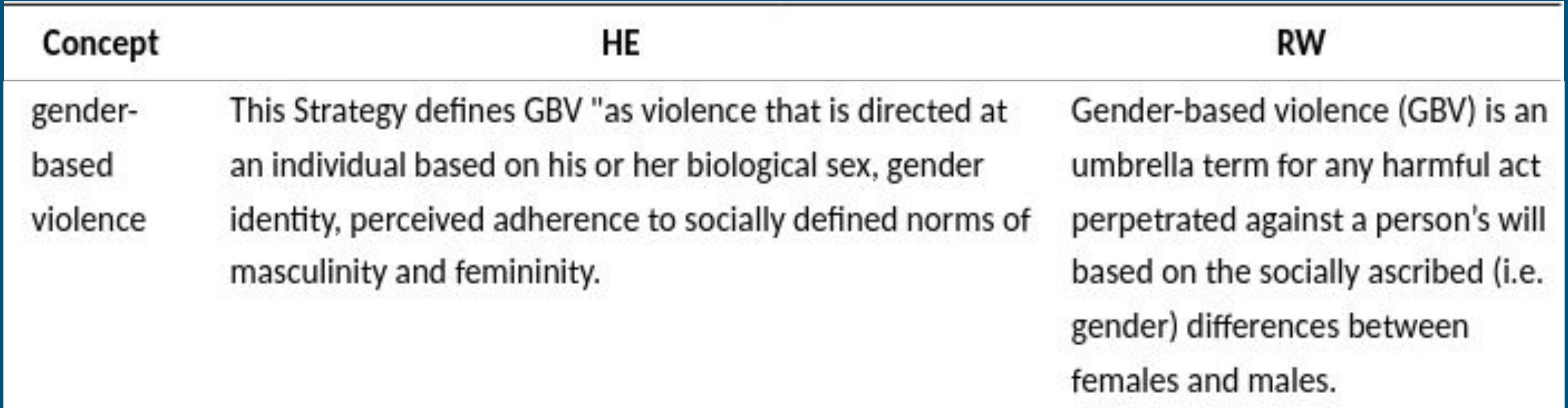
| Concept | Shared | HE | RW |
|---|---|---|---|
| humanitarian reform | (0/9) | challenge, development [recent change], matter | module, initiative, issue, priority, reform, solution |
| sustainability | criterion, goal, indicator, issue, principle, theme, topic (7/25) | category, cornerstone, driver, 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 |

# Definitional contexts



| 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. |

# Definitional contexts



| Concept | HE | RW |
|---|---|---|
| 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. |