Aggregating and Visualizing Collocation Data for Humanitarian Concepts (Short Paper)

Loryn Isaacs[1], Pilar León-Araúz[1]

[1]*University of Granada, C/ Puentezuelas, 55, Granada, Spain*

Abstract

Analyzing a term's collocations often offers insight into domain-specific usage, yet manually comparing large data sets of collocations can be unfeasible. This necessitates programmatic techniques that aggregate large quantities of collocation data and condense results into manageable visualizations. This paper presents a method to quickly process hundreds of thousands of corpus queries with a combination of the Sketch Engine API and related open-source software. A preliminary web application is offered to explore aggregated collocation data for the humanitarian concepts that make up the Humanitarian Encyclopedia. Potential applications are discussed with regards to the study of conceptual variation in the humanitarian sector.

Keywords[1]

Humanitarian terminology, collocation, visualization, corpus, Sketch Engine API

1. Introduction

Collocations are a well-known source of information regarding a term's meaning and usage in a specialized context. The analysis of co-occurring lexical items has been formalized in corpus management systems, including via user interfaces that present summaries of collocational behavior and allow for further data exploration. Sketch Engine's word sketch feature is one example, offering a summary of how strongly and frequently a term is associated with various types of collocates [1]. This tool has been utilized to conduct concept analyses for the Humanitarian Encyclopedia, an open platform for linguistic data and expert discussion on humanitarian terminology [2]. To better facilitate the development of the encyclopedia's concept entries, a data exploration method was developed to condense collocation data from many queries into an interactive visualization. This paper summarizes the workflow used to extract and explore bulk frequency data with the Sketch Engine API.

The following sections overview the Humanitarian Encyclopedia and its corpus of domain-specific texts (Section 2), the API-based data collection method using Python (Section 3), the Flask web application designed to explore the data set (Section 4), and areas for future research (Section 5).

2. The Humanitarian Encyclopedia

The Humanitarian Encyclopedia is an ongoing collaborative project from the Geneva Centre of Humanitarian Studies that focuses on studying conceptual variation and elucidating an internationally shared understanding of key humanitarian concepts. It aims at defining and documenting the dynamics of concepts that are particularly controversial, fuzzy, or ill-defined within the humanitarian action domain. It currently focuses on 129 concepts, including HUMANITARIANISM itself, as well as a range of

[1]2nd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2023, June 29–30, 2023, Lisbon, Portugal

EMAIL: lisaacs@ugr.es (L. Isaacs); pleon@ugr.es (P. León-Araúz)

ORCID: 0000-0003-0267-4853 (L. Isaacs); 0000-0002-8520-2749 (P. León-Araúz)

events, strategies, entities, and other phenomena related to humanitarian activities (FOOD SECURITY, DO NO HARM, INDEPENDENCE, etc.).

Each concept entry is created according to an approach that combines corpus-driven knowledge provided by terminologists with expert knowledge from humanitarian practitioners or academics. Each entry offers a blend of quantitative and qualitative data that describe a concept's primary characteristics, the degree to which its usage is homogeneous, and debates among humanitarian actors as to its meaning or institutional value (see the example for DO NO HARM in **Figure 1**). The technologies and procedures used to generate content for the encyclopedia's concept entries are summarized in [3].

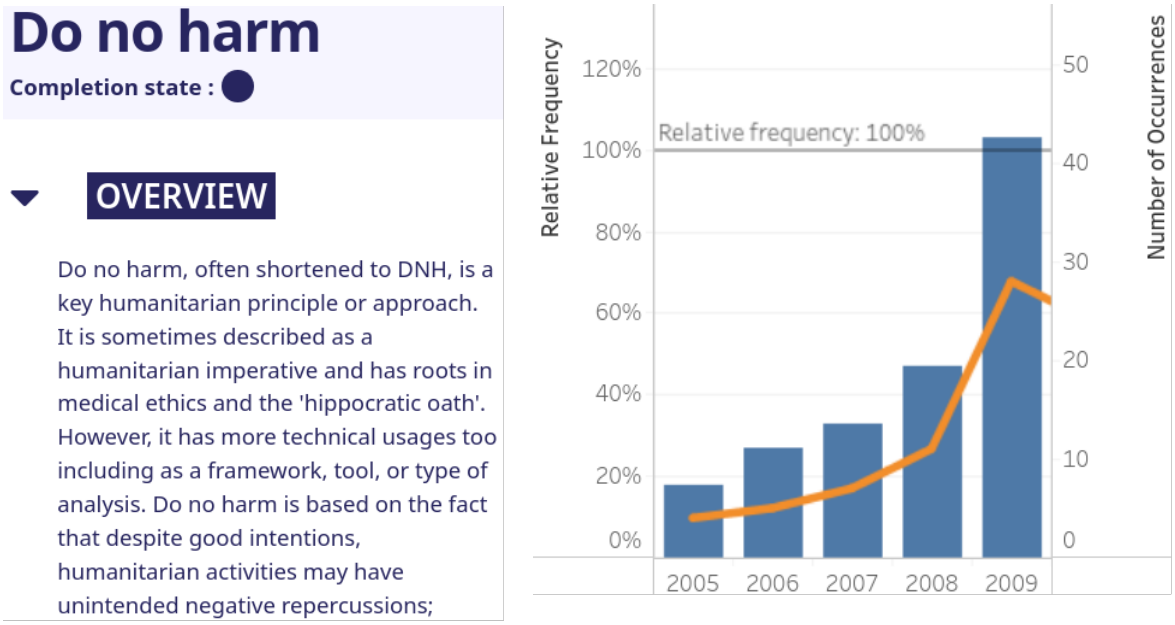


Figure 1: Humanitarian Encyclopedia entry overview and visualization

Linguistic data for concept entries are extracted from the Humanitarian Encyclopedia corpus, which consists of texts compiled from humanitarian websites and public databases, such as UN Women[2] and ReliefWeb[3]. The corpus contains 71 million words and is comprised of annual reports, strategy documents, and general documents in English published from 2005 to 2019. Authors include a variety of organizations associated with humanitarian efforts, from international actors to local groups. These are classified into 26 organization subtypes, with a majority representing nongovernmental organizations, the United Nations, intergovernmental organizations, and the Red Cross. Its 4,814 documents are tagged by region, originating mostly from Europe, North America, and Asia, year of publication, document type, and organization type.

For the Humanitarian Encyclopedia, the phenomenon of collocation plays a key role in identifying the semantic content of a concept. This can include identifying hypernyms, hyponyms, causes, effects, term variants and antonyms, as well as controversies, often in reference to a concept's shared definition (or lack thereof) and its implementation in humanitarian response. These units tend to be extracted in Sketch Engine through statistical association with an emphasis on nouns, adjectives, and verbs. Extraction may be based on collocational strength with the logDice score [4], the identification of multiword terms, or the analysis of relevant semantic relations (most often hyponomy, meronymy, and causality) [5]. For instance, collocations of EPIDEMIC allow accessing its multiple conceptual dimensions by classifying the multiword terms in which epidemic is the head: pathogen (e.g., *HIV epidemic, Zika virus epidemic*), cause (e.g., *obesity epidemic, tobacco epidemic*), morbidity (*global epidemic, localized epidemic*), time (e.g., *recurrent epidemic, seasonal epidemic*), severity (e.g., *deadly epidemic, severe epidemic, acute epidemic*), etc. Other (more distant) collocations highlight the most mentioned countries struck by epidemics. Verbs, in turn, point to the effects of epidemics or the actions that can be undertaken before, during, and after an epidemic outbreak. Most verbs occurring with

[2]https://www.unwomen.org/en
[3]https://reliefweb.int/

epidemic as a subject indicate its sudden and violent nature, as they are impact-related verbs (*hit, strike, rage, sweep, break out, devastate*), whereas most verbs occurring with epidemic as an object are response-related (*contain, reverse, avert, combat, fight, prevent, curb, stop*, etc.), indicating a subsequent phase in the event of an epidemic. More rarely, there are also verbs indicating anticipation (*prevent, avert, detect*).

Collocations also help in identifying knowledge rich contexts, such as "Every year, millions of people around the world experience the devastating effects of *disasters such as floods, droughts and epidemics*" or "Health ministers and other experts from the region exchanged views on strengthening the resilience of health-care systems against *epidemics, armed conflicts and other emergencies"*. The analysis of large collections of KRCs shows that depending on how EPIDEMIC is categorized (e.g., disaster, emergency), it has slightly different clusters of sibling concepts. For example, when categorized as disaster, threat, or calamity, sibling concepts are mostly related to natural hazards, whereas when categorized as emergency, shock, or factor, epidemic is part of a larger humanitarian frame, including siblings such as *food insecurity*, *nutritional crises*, *conflicts*, *displacement*, *political crises*, *poverty*, etc.

All of these collocation-related analyses point to a dynamic conceptualization that can also be correlated with corpus metadata (i.e., organization type, publication date, etc.), which makes the foundations of conceptual variation analysis. For instance, IGOs show more collocates related to epidemic types (*SARS, Marburg, dengue, polio, hepatitis, influenza*) or their impact/origin-related attributes (*deadly, waterborne, devastating*), whereas NGOs focus on attributes (*devastating, deadly, lethal*) but especially impact (*rage, break, hit*) and response-related verbs (*contain, prevent, reverse*).

For the development of the Humanitarian Encyclopedia, data derived from the above methods are visualized in each concept analysis as a way to interpret data, correlate variables, and transfer knowledge to humanitarian experts. Plots can include standard visualizations, such as histograms of text types and maps of document source countries, as well as bespoke visualizations when merited by a unique linguistic phenomenon. While this modus operandi is necessary and helpful for conducting individual analyses, a contrastive approach could offer a more global perspective on the relationships between terms, collocations, and communicative contexts. Such an approach, however, is not practical without automating queries and developing a convenient means to explore data. In response, the following method was developed to visualize the entirety of the Humanitarian Encyclopedia's collocation data in one resource.

3. API-based data collection

Collecting collocation data for each of the Humanitarian Encyclopedia's concepts required utilizing Sketch Engine's API, which allows developers to programmatically execute queries. The software employed to manage API calls was Sketch Grammar Explorer (SGEX), an API wrapper written in Python [6]. This tool was employed in conjunction with NoSketch Engine, the software's open-source variant [7], in the form of a Docker container maintained by the Eötvös Loránd University Department of Digital Humanities [8]. Together they provided a means to locally query the corpus and store results as a single data set.

The primary benefit of using a local instance of NoSketch Engine's API, as opposed to manual file download or Sketch Engine's rate-limited API, is to substantially increase data collection rates for large numbers of queries. The present data collection task, as described below, would have required over 5 months of continual operation to execute API requests to Sketch Engine's main server at the allowed rate. In contrast, making requests locally on a consumer desktop with a recent Core i5 Intel processor reduced this figure to under 10 hours.

To allow for more granular data manipulation, collocation frequency data were collected by combinations of text types. While previous visualizations, such as in **Figure 1**, summarize data for single concepts and text types (e.g., occurrences of DO NO HARM by year or by region or by organization type), the current method accepts multiple restrictions. Users could, for example, select multiple concepts at once by year and by region and by organization type (e.g., occurrences of INDEPENDENCE and IMPARTIALITY for 2013-2015 in European NGOs. To do so, this required making 2,316 API calls for each concept, or 298,764 in total, although such granular text type constraints returned no

concordances in almost half (135,929) of these calls. When API calls did retrieve hits, up to 20 of the top collocations by logDice were included. Where possible, CQL rules incorporated common abbreviations and variations. Example query syntax and results with combinations of text type restrictions are shown in (1) and Table 1.

```
(([ lemma_lc = "gender" ] [ word = "(B|b)ased" ] | [ lemma_lc = "gender-based" ])[
lemma = "violence" ]) | [ lemma_lc = "GBV" ] within < class ( DATE = "2004-
2005|2005" ) & ( REGION = "Europe" ) & ( TYPE = "General_Document" )/>
```
(1)

Table 1
Random sample of collocation data with text type restrictions

Row #	Collocate	Concept	logDice	Freq	Date	Region	Org type	Doc type
585041	negative	impact	5.96	10	2018	Asia	NGO	
996561	environmental	protection	2.36	3	2008		Found[b]	AR[a]
785890	standardized	monitoring	7.44	4	2019	Europe	IGO	
1108623	hygiene	sanitation	6.09	40	2017	MENA		AR[a]
502451	crisis	funding	1.12	3	2018	MENA		AR[a]

[a]Activity report
[b]Foundations/funds

To prepare the data for visualization, API responses were merged into a tabular format and cleaned, including the removal of unwanted collocates, such as non-words, non-Latin characters, and auxiliary verbs. This was done in two passes, first by identifying unwanted strings automatically with the *unicodedata* Python package (**Appendix A**), and then manually curating an exclusion list of remaining unwanted items (e.g., *xiii, the*, *be, will*). The final data set amounted to 1,236,194 rows with 19,812 unique collocates disaggregated by text type. Among these collocates the most frequent was *humanitarian*, at 6,947 cases, followed by *disaster*, *health*, *community*, and *development*. To ensure the veracity of automatically retrieved results, a sample was compared with those retrieved manually via user interface.

4. Web application design and purpose

A tool was built with Python and the Dash web application framework [9] to visualize the prepared Sketch Engine API data. It consists of four elements: an interactive scatter plot, a series of dropdown and slider components for adjusting parameters, a table of summary statistics, and a URL generator that redirects user-selected data points to Sketch Engine. These elements are generated automatically based on the shape of the data set, e.g., adding text filters for corpus structures and numeric filters for frequencies and logDice scores. Visualizations are generated as users apply filters, offering a standardized means for analysis and evaluation tasks. The data can be further restricted by sample size, to show the top *n* collocates, as well as be displayed with faceting, to show data subsets in separate plots. Users can then identify areas of interest that could previously have been cumbersome to explore with disparate queries.

One task facilitated by the visualization is assessing conceptual variation across a specialized domain. For example, **Figure 2** shows a selection of the top collocates for the humanitarian concept of ACCESS by three organization types: IGO, NGO, and Red Cross. Here an area of interest is the relationship between the subjects and objects for which access is a challenge. While *sanitation* appears in each of the three organization types, *internet* and *energy* are exclusive to IGO and *education* is exclusive to NGO. *Detainee,* a type of population that both requires access to resources and which organizations seek access to, is exclusive to Red Cross. These differences may indicate how parties represented in the corpus focus their activities on discrete objectives and populations. Discussion of the meaning of ACCESS, then, could consider commonalities and differences measurable in the corpus

regarding semantic roles: party demanding access, population needing access, authority granting access, and service being accessed.

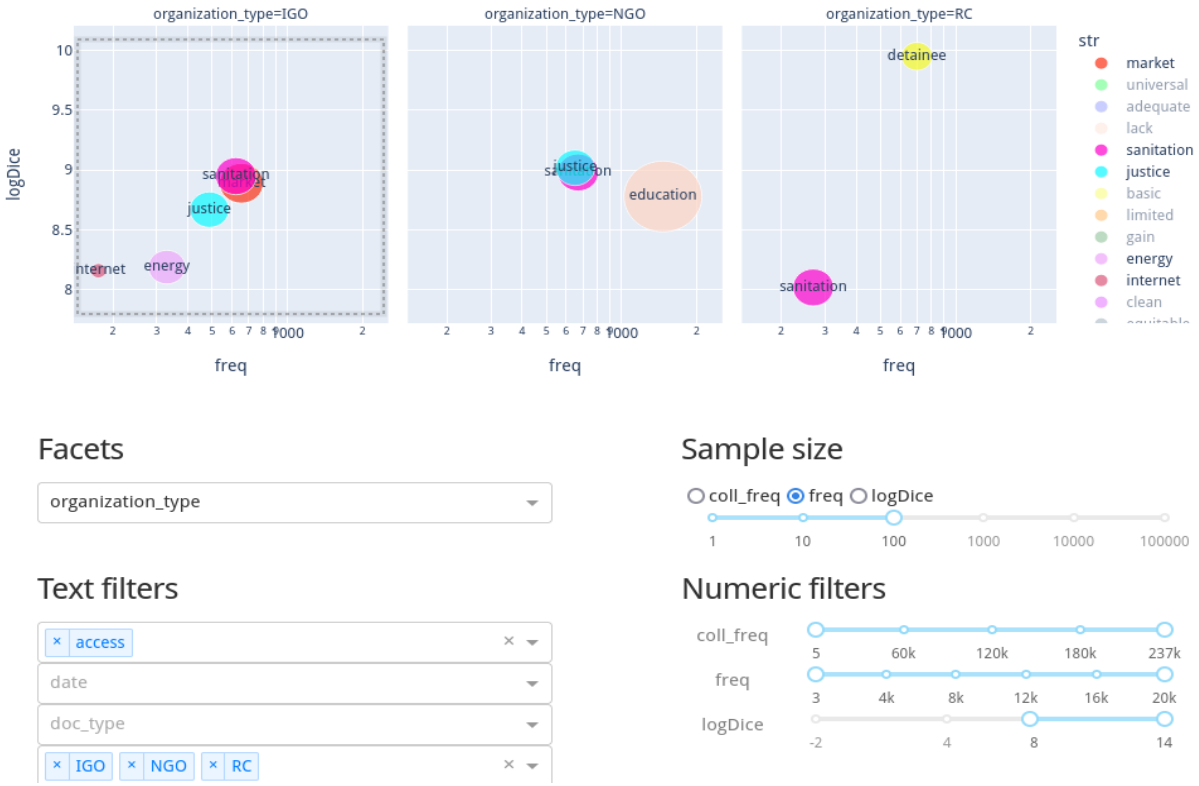


Figure 2: Top collocates for ACCESS by organization type

As seen in **Figure 2** and other examples provided as appendices, the discovery of possible correlations could aid terminological analysis and help transmit results to humanitarian experts. In that regard, one consideration is the need to provide sufficient contextual information for guiding proper data interpretation. Supplementary visualizations describing the shape of the data set and its limitations would be beneficial for users. For instance, among the 129 concepts there is a wide range of frequencies. While *development* and *community* each appear as terms (as opposed to collocates for other concepts) over 23,000 times, several polylexical terms have very few cases, like *humanitarian-development nexus* (99) and *humanitarian imperative* (132). Evaluating how combinations of text type restrictions influence the composition of results will be a necessary next step.

5. Future applications

The data management approach described here addresses a need for the Humanitarian Encyclopedia to streamline corpus-based analysis of humanitarian concepts and their collocations. It is an example of an open-source method for integrating the Sketch Engine API into a workflow for terminological research. The increased rate and scale of data extraction encourages more techniques for the exploration and analysis of specialized corpora. The immediate interest for the Humanitarian Encyclopedia will be to research how key terms behave across the humanitarian sector, particularly their degree of standardization among actors and the prevalence of controversies.

The prototype interface described in this article is part of a larger effort to create an open-source dashboard for visualizing the Humanitarian Encyclopedia corpus. A central aim is to track developments in the current usage of humanitarian concepts across the sector. This will be aided by developing a means to visualize data from multiple sources and integrating other query systems in addition to Sketch Engine's. While the automation of many corpus queries allowed for the creation of a new data set, with automation comes additional challenges for presentation and contextualization.

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).

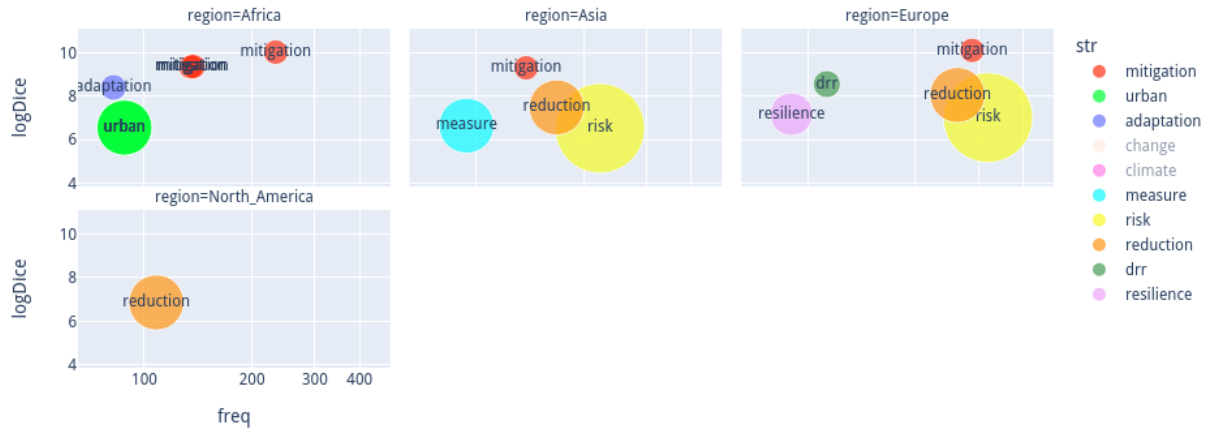
7. References

[1] A. Kilgarriff, V. Baisa, J. Bušta *et al.* The Sketch Engine: Ten years on. Lexicography ASIALEX 1, 7–36 (2014). doi:10.1007/s40607-014-0009-9.
[2] Humanitarian Encyclopedia. URL: https://humanitarianencyclopedia.org.
[3] S. Chambó, P. León-Araúz, Visualising lexical data for a corpus-driven encyclopaedia, in: I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (Eds.), Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference, Lexical Computing, Brno, Czech Republic, 2021, pp. 29-55.
[4] P. Rychlý, A lexicographer-friendly association score, in: P. Sojka, A. Horák (Eds.), Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008, Masaryk University, Brno, Czech Republic, 2008, pp. 6–9.
[5] P. León-Araúz, A. San Martín, P. Faber, Pattern-based word sketches for the extraction of semantic relations, in: P. Drouin, N. Grabar, T. Hamon, K. Kageura, K. Takeuchi (Eds.), Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 73–82.
[6] L. Isaacs, Sketch Grammar Explorer. doi:10.5281/zenodo.6812335.
[7] P. Rychlỳ, Manatee/Bonito-A modular corpus manager., in: P. Sojka, A. Horák (Eds.), First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007, Masaryk University, Brno, Czech Republic, 2007, pp. 65–70.
[8] Eötvös Loránd University Department of Digital Humanities, NoSketch-Engine-Docker. URL: https://github.com/ELTE-DH/NoSketch-Engine-Docker
[9] P. T. Inc., Collaborative data science, 2015. URL: https://plot.ly

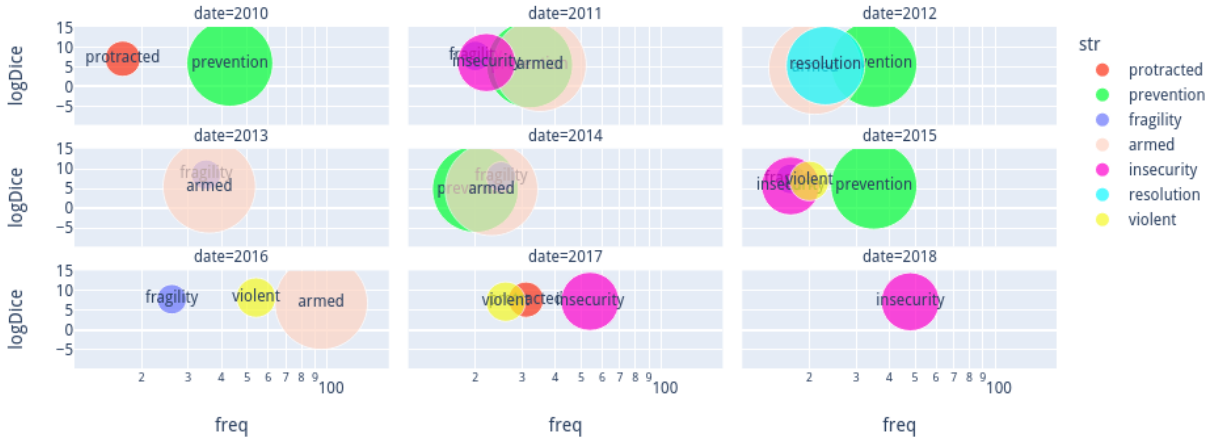
8. Appendices

Appendix A: Function for automatic string exclusion

```
for collocate in list_of_unique_strings:
    normalized = unicodedata.normalize('NFD', collocate)
    canonical = u"".join([
        char for char in normalized
        if not unicodedata.combining(char)])
    regex_drops.update(
        re.findall(
            re.compile(".*[^a-zA-Z\-\.\']+.*", re.UNICODE), canonical))
```



Appendix B: Top collocates for ADAPTATION by region (excluding *climate change*)



Appendix C: Top collocates for CONFLICT by *date* in European documents