Visualising Lexical Data for a Corpus-Driven Encyclopaedia

Santiago Chambó[1], Pilar León-Araúz[2]

[1, 2] Department of Translation and Interpreting,
University of Granada, Buensuceso, 11, 18002 Granada (Spain)
E-mail: santiagochambo@ugr.es, pleon@ugr.es

Abstract

The Humanitarian Encyclopedia (HE) is an ongoing corpus-driven project that aims at defining and documenting the dynamics of 129 concepts that are particularly controversial, fuzzy or ill-defined within the humanitarian action domain, thus enhancing communication in a sensitive area. In the HE, each entry is created according to an approach that combines corpus-driven knowledge with expert knowledge. Concept entries are authored by field experts who are provided with a Linguistic Analysis Report (LAR) created by a team of linguists. In LARs, HE linguists support their claims by i) presenting, quantifying and categorising textual data and by ii) making comparisons among subcorpora, which are created based on the corpus metadata (i.e. document type, region, organisation type, publication year). This article presents the visualisations created by HE linguists to represent both semantic information (i.e., conceptual combinations and non-hierarchically related concepts) and quantifiable concordance and collocational data. This includes approaches to disaggregating measures according to different kinds of subcorpus types and strategies to represent collocational intersections among subcorpora (i.e., collocates occurring in multiple subcorpora) as well as collocates unique to each subcorpus. Other concept-specific visualisations were also designed and are examined in this article.

Keywords: lexical data; visualisation; concept

1. The Humanitarian Encyclopedia: a Corpus-Driven Project with Lexical Data Visualisations

The Humanitarian Encyclopedia (HE; https://humanitarianencyclopedia.org/home) is an ongoing corpus-driven project that aims at defining and documenting the dynamics of 129 concepts that are particularly controversial, fuzzy or ill-defined within the humanitarian action domain. In the humanitarian domain there are many stakeholders (i.e. academics, practitioners, decision-makers) who do not always share a consensual understanding of humanitarian concepts, such as VULNERABILITY, RESILIENCE or AID DEPENDENCE. Although, at least theoretically, they all share common principles and values, even the very notion of HUMANITARIANISM raises controversial issues. The humanitarian sector is thus a highly dynamic domain due to different factors, such as history, academic and professional disciplines, culture, religion, organisational cultures and contexts, which

are the reasons behind both its richness and controversies. Conceptual controversies raise operational, political, societal and educational challenges that can hinder the effectiveness of humanitarian action in a global world.

In this context, the initiative of the HE aims to cover an existing gap in the humanitarian sector contributing to the public good. As acknowledged on its website there is a current need for "creating a common understanding and formulation of the key humanitarian concepts to build bridges and promote an open dialogue to improve collective humanitarian action".

In the HE, each entry is created according to an approach that combines corpus-driven knowledge with expert knowledge. Concept entries are authored by field experts who are provided with a Linguistic Analysis Report (LAR) stored in the Linguistic Analysis Portal for the Humanitarian Encyclopedia (https://sites.google.com/view/humanitarianencyclopedia). A team of linguists is in charge of producing LARs for each concept based on data extracted from a corpus of humanitarian texts. Every LAR provides an overview of how a concept is understood explicitly and implicitly in humanitarian discourse and proposes a definitional template for it. Each LAR is generally composed of the following elements:

- Frequencies, which allow experts to see the regions, document types, years and organisation types where the concepts appear more relevant.

- Definitions, whether standardised and authoritative (if found in the corpus), or *ad hoc* (based on implicit categorisation), together with a summary of definitional elements and a comparison based on corpus metadata.

- Related concepts: indicating how concepts change their relational behaviour based on organisation type, geographical regions or time (e.g. causes and consequences, affected population, subtypes classified on different conceptual dimensions, ways of managing humanitarian concepts, etc.).

- Frequent collocations, mostly nouns, adjectives and verbs, showing other surrounding concepts in the corpus, which allow experts to understand the different facets of the concept over time and across organisations.

- Synonyms and antonyms, where applicable, together with the sources from which they were extracted.

- Usage over time, where applicable, according to both the HE corpus and Google Ngram Viewer.

- Trends, debates and controversies surrounding each concept, which is one of the richest elements and requires extensive manual curation.

HE linguists decided to include visualisations to aid their own analyses and make lexical data more accessible and thought-provoking for HE authors, which, due to space limitations, is the focus of this paper.

Projects driven by lexical data require visualisation strategies that facilitate data interpretation and enable knowledge transfer (Allen, 2017). Firstly, making sense of any kind of data without the support of graphical representations constitutes a cognitively challenging task, and linguistic data is no different (Siirtola et al., 2010). Secondly, in a multidisciplinary project where linguists and field experts interact, the visualisation of lexical data serves as an intermediary between both stakeholders.

The remainder of this paper is structured as follows. Section 2 describes the materials and methods used by the HE linguists. Section 3 presents the visualisations created to support lexical data interpretation. In Section 4 conclusions and future lines of research are presented.

2. Materials and Methods

This section describes the materials and methods used to create datasets of lexical data and to build visualisations based on such datasets.

2.1 Materials

2.1.1 Sketch Engine

Sketch Engine (www.sketchengine.eu) is a browser-based software that enables users to build, analyse and query corpora (Kilgarriff et al., 2004). It contains many tools and functionalities that can be combined. Table 1 provides a summary of the main tools and functionalities used for the purposes of this work.

Tool	**Description**	**Functionality**	**Description**
Concordance	Queries a corpus and return results in context, which can be sorted, filtered and processed with many additional functionalities. Complex searches are conducted with CQL[1].	Hide Sub-Hits filter	Removes sub-hits from matches obtained with queries containing ranges (e.g., {1,3}), only keeping the longer results.
		Frequency	Computes frequencies from results, generating frequency reports.
		Collocations	Compute collocations from results.
Word Sketch	Provides a summary of a search term's collocates and other surrounding words. Results categorised by grammar relations defined by a file containing a set of rules known as sketch grammar.	-	-

Table 1: Main tools and functionalities used in Sketch Engine.

2.1.2 The HE Corpus

The HE Corpus is a collection of 4,824 humanitarian documents published between 2004 and 2019, which amount to a total of 84,926,707 tokens and 71,201,157 words. Documents are tagged with metadata according to the type and subtype of issuing organisation, region, year of publication and document type. These are referred to in Sketch Engine as text types. Table 2 contains all metadata fields and values associated with each document save for organisation subtype because it is not used in the visualisations described in this paper.

The corpus was uploaded onto Sketch Engine and processed with a custom sketch grammar that combines Sketch Engine's default sketch grammar for English with

[1] Corpus Query Language (CQL), as referred to in Sketch Engine documentation, is a concordance notation that allows users to search corpora for complex grammatical and lexical patterns. It is based to a large extent on the Corpus Query Processor language (or QQP-syntax) implemented in Corpus Workbench and developed by Christ et. al (1999).

the EcoLexicon Semantic Sketch Grammar (León-Araúz & San Martín, 2018) and an unpublished set of rules for multi-word term extraction (see Section 4.5).

Text Types	**Classes**
Organisation Type	**NGO** (Non-Government Organisations), **NGO_Fed** (NGO Federations), **IGO** (Intergovernmental Organisations), **RC** (Red Cross/Crescent), **Net** (Networks), **Found** (Foundations/Funds), **State** (Government/State Entities), **RE** (Religious Entities), **C/B**, Project and WHS
Region	**Africa**, **Asia**, **CCSA** (Caribbean, Central and South America), **MENA** (Middle East and North Africa), **North_America**, **Oceania**
Year	Between **2004** and **2019**
Document Type	**General_Document**, **Activity_Report**, **Strategy**

Table 2: Pertinent text types in the HE Corpus

2.1.3 Tableau

Tableau (www.tableau.com) is a commercial data visualisation software, which has been used in previous projects to visualise linguistic data (Allen, 2017; Desagulier, 2019). It interprets datasets in multiple formats and provides the user with a graphic interface that enables him or her to create visualisations by combining a wide range of options. The visualisations described in this paper were created with Tableau Desktop. To embed our visualisations on the website where the LARs are published, each visualisation has to be uploaded onto a Tableau Public profile (https://public.tableau.com/).

2.1.4 Google Data Studio

Google Data Studio (datastudio.google.com) is a browser-based visualisation solution similar to Tableau. It is solely used to create filtrable and searchable tables (see Section 4.7) because Tableau does not offer such visualisation option.

2.1.5 Spreadsheet software

To create datasets in supported formats, we used Microsoft Excel for the visualisations built with Tableau, and Google Sheets for Google Data Studio.

2.2 Methods

This subsection provides a brief overview of the methods used to extract data from the HE Corpus with Sketch Engine and to create the datasets in a way that can be interpreted correctly by Tableau. For clarity, specific steps and procedures for each visualisation are described in Section 4.

Data is extracted from the HE Corpus through two methods with the Sketch Engine querying functionalities. The first method entails querying the corpus with CQL expressions by using the Concordance tool and its processing options (see Table 1 in Section 3.1.1). With this method, we aim at creating datasets that contain string value fields for lexical units and associated measures. This method also enables us to conduct restricted searches in specific portions of the corpus (i.e., subcorpora) by specifying document metadata in our CQL queries. A second method uses the Word Sketch functionality to query the corpus. Data is therefore collected from specific grammatical relation reports.

Data Fields	**Data Type in Tableau**	**Description**
Lexical units	Dimension (string)	Any word or words extracted by querying the corpus (e.g., concordance matches, multi-word expressions, collocates, contexts, etc.) to be displayed in a visualisation.
Organisation type	Dimension (string)	Metadata values from documents in the corpus
Year	Date	
Document type	Dimension (string)	
Region	Dimension (string)	
Frequency (absolute frequency)	Measure (whole number)	Number of occurrences in the corpus
Relative frequency	Measure (decimal percentage)	Subcorpus frequency divided by the frequency of a query in the entire corpus; expressed as a percentage (Kilgarriff et al., 2015)
logDice	Measure (decimal number)	Score expressing typicality of extracted collocations; independent of corpus size and recommended to compare phenomena among subcorpora (Rychlý, 2008)

Table 4: Data fields used to build the visualisations

In Sketch Engine, each query generates a report that we treat with spreadsheet software to create CSV files with a data structure that can be processed by Tableau. These datasets are built by processing data fields and values from reports for single queries obtained from Sketch Engine, as well as combining results from multiple reports. Table 4 details all the data fields sourced from Sketch Engine reports and used to build datasets.

3. Visualisations

This section presents the visualisations created to support lexical data interpretation in LARs. Each subsection is organised around the datasets used to build each visualisation. Visualisations built with the same dataset are discussed in the same subsection. Unfortunately, due to length constraints, we will not provide detailed instructions of how each visualisation was built on Tableau.

By default, all LARs contain at least six visualisations, namely:

- a frequency histogram, disaggregating frequency by year of publication, organisation type, region and document type;
- a map, representing absolute frequency and relative frequency by region;
- a collocation histogram, showing the collocates by year with the highest logDice score;
- a dual axis bar and line chart, representing relative frequency and absolute frequency by year, region, organisation type and document type;
- a unique collocate packed bubble chart, representing collocates unique to each organisation type and their logDice scores; and
- a bar chart, representing collocates shared by more than two organisation types.

Additional visualisations are created depending on the nature of each concept entry. This article also covers the following *ad hoc* visualisations:

- square treemaps, detailing conceptual combinations and coordinated concepts;
- a histogram, representing manually curated contexts to represent conceptual development across time; and
- sortable and searchable tables containing manually curated contexts.

3.1 Frequency Histogram

A frequency histogram represents the frequency of a search term disaggregated by year of publication. This only requires a simple dataset that can be easily obtained from Sketch Engine. With it, our histogram can also disaggregate yearly frequencies by organisation type, region and document type.

To begin, we query the corpus with the Concordance tool by using the CQL expression [lemma_lc="x"] where x is the term or list of terms designating a concept. We then use the Frequency functionality to compute the frequencies of the search words in the concordance lines. Lastly, we select the Line Details pre-set, which generates a report detailing every document in the corpus that contains the search term. Each record represents a document and details all its text type metadata, frequency and a percentage of the total concordances (see Figure 1). This report is exported as a CSV file.

	Class.DATE	Class.ORGANIZATION_SUBTYPE	Class.TYPE	Class.REGION	Class.ORGANIZATION_TYPE	Class.ID	Frequency ↓	% Of conc.	
1	2013	IFRC	General_Document	Europe	RC	GD-101	21	5.92 %	···
2	2014	NGO_Nat	General_Document	Europe	NGO	GD-255	16	4.51 %	···
3	2005	0	General_Document	Europe	C/B	GD-36	15	4.23 %	···
4	2013	RCNS	Activity_Report	Asia	RC	AR-3568	12	3.38 %	···
5	2014	IFRC	General_Document	Europe	RC	GD-102	11	3.10 %	···
6	2015	IFRC	General_Document	Europe	RC	GD-99	8	2.25 %	···
7	2016	NGO_Int	Activity_Report	North_America	NGO	AR-2006	8	2.25 %	···
8	2018	IFRC	General_Document	Europe	RC	GD-137	8	2.25 %	···
9	2011	0	General_Document	Europe	C/B	GD-56	7	1.97 %	···
10	2007	0	General_Document	Europe	C/B	GD-45	7	1.97 %	···
11	2013	UO	Activity_Report	Europe	NGO_Fed	AR-2675	7	1.97 %	···
12	2008	0	General_Document	Europe	C/B	GD-46	6	1.69 %	···
13	2012	NGO_Nat	General_Document	Europe	NGO	GD-251	6	1.69 %	···

Figure 1: Frequency Line Details report for LEAVE NO ONE BEHIND

The resulting raw CSV file requires minimal treatment with spreadsheet software because the target data structure mirrors the one generated by Sketch Engine, as can be seen in Figure 1. This means creating a spreadsheet with each row representing a document, six columns containing text type metadata and a seventh column containing frequency values. The percentage of total concordances is discarded. After treatment, the CSV file is ready to be added on Tableau as a data source.

In Tableau, fields for text type metadata are set as dimensions, whereas frequency is set as a measure. Figure 2 shows the default view of our frequency histogram as published in the LAR for LEAVE NO ONE BEHIND. On the right there are three toggle options that allow users to further disaggregate frequencies by increasing the number of axes.

Field experts can thus observe that LEAVE NO ONE BEHIND appears mostly in documents published in Europe, followed by North America. Overall, the top five contributors in terms of occurrences are IGO, NGO, NGO_Fed, Net and State

organisations. IGO documents generate more than half of all occurrences in the HE Corpus. Contributions from other organisation types are significantly smaller.

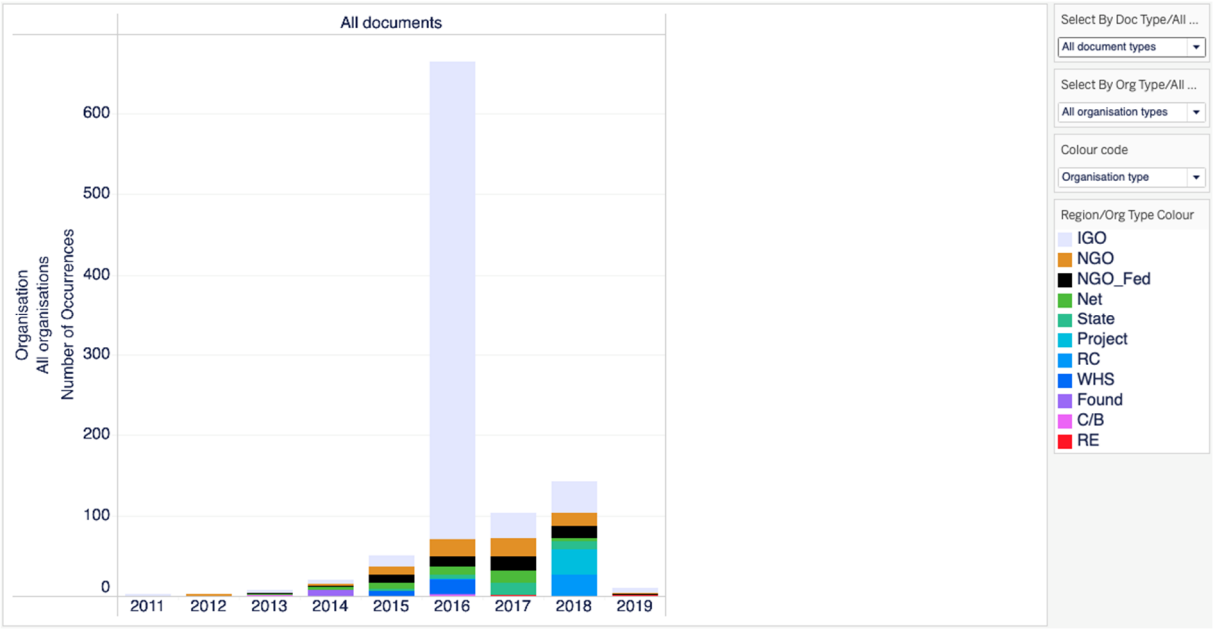


Figure 2: Default view of the frequency histogram for LEAVE NO ONE BEHIND

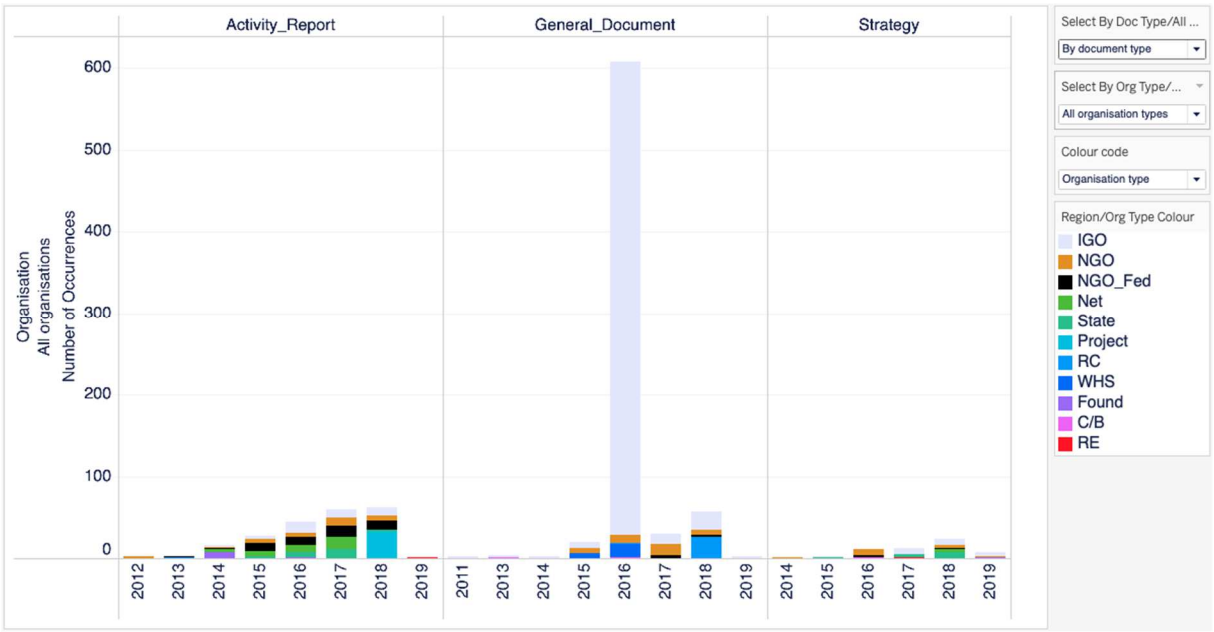


Figure 3: A dynamic axis view disaggregating yearly frequencies by document type.

3.2 Map and Relative Frequency Bar Charts

Comparing absolute frequency and relative frequencies can be achieved by building a dataset for each concept. It can be used to create two visualisations. The first is a map representing absolute and relative frequencies by region. The second constitutes a set of bar charts that focus on comparing absolute and relative frequencies disaggregated by year, organisation type, region and document type.

As with the dataset described in Section 4.1, we queried the corpus with the Concordance tool by using CQL. In the Frequency functionality, we used instead the Text Types pre-set, which generates a report detailing the absolute frequency, relative frequency and percentage of total concordances for each text type in the corpus. Figure 4 shows part of this report with values for organisation types and subtypes.

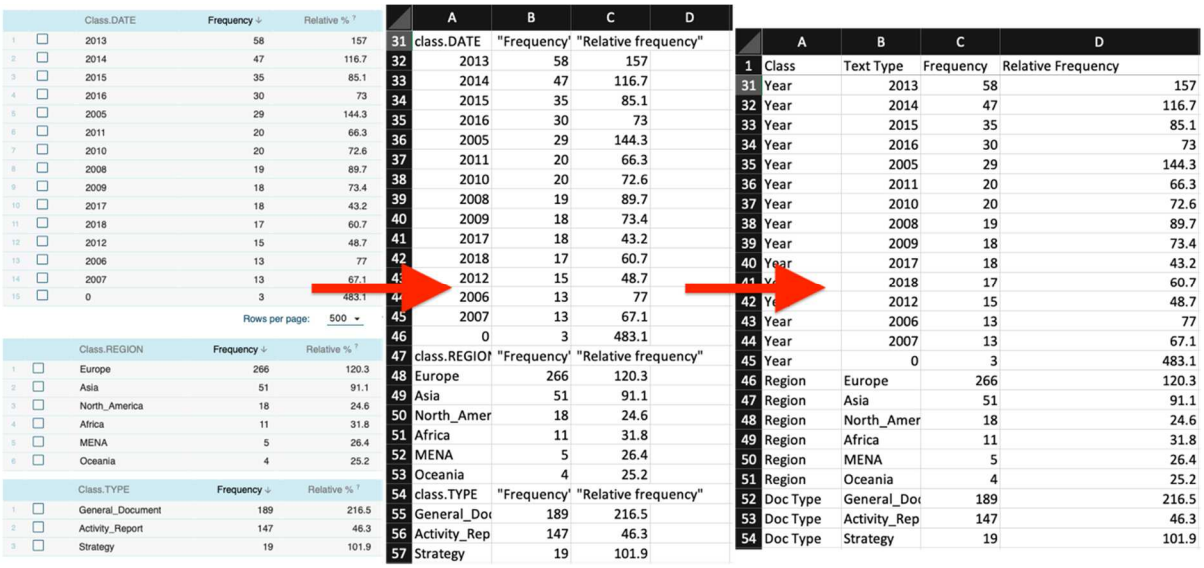
Text type reports as CSV files require more treatment with spreadsheet software. Each record in the report corresponds to a kind of text type, and this is not specified in the raw CSV file. This means that all text types are contained in the same column. For this reason, a new column has to be added to disambiguate text types. Figure 5 illustrates the spreadsheet treatment process to obtain a data structure that can be interpreted correctly by Tableau.

		Class.ORGANIZATION_TYPE	Frequency ↓	Relative % ?	% Of conc.
1	☐	NGO	95	72	26.76 %
2	☐	RC	93	181.9	26.20 %
3	☐	C/B	81	706.1	22.82 %
4	☐	NGO_Fed	45	80.2	12.68 %
5	☐	Net	15	79.6	4.23 %
6	☐	WHS	10	793.2	2.82 %
7	☐	Found	6	43.9	1.69 %
8	☐	IGO	4	3.9	1.13 %
9	☐	State	4	15.4	1.13 %
10	☐	RE	1	12.1	0.28 %
11	☐	Project	1	45.7	0.28 %

Rows per page: 500 ▾ 1–11 of 11 |<

		Class.ORGANIZATION_SUBTYPE	Frequency ↓	Relative % ?	% Of conc.
1	☐	0	105	241.5	29.58 %
2	☐	IFRC	56	950.9	15.77 %
3	☐	NGO_Int	48	60.1	13.52 %
4	☐	NGO_Nat	42	134.8	11.83 %
5	☐	RCNS	37	175.6	10.42 %
6	☐	UO	35	189.1	9.86 %
7	☐	Net_GP	8	199.2	2.25 %
8	☐	NGO_Fed_NA	7	25.2	1.97 %
9	☐	NGO_Reg	5	27.9	1.41 %
10	☐	UN_OPA	4	7.1	1.13 %
11	☐	AA	4	19.5	1.13 %
12	☐	NGO_Int_NO	3	489	0.85 %
13	☐	NGO_Nat_Net	1	37.1	0.28 %

Figure 4: Text type report for HUMANITARIANISM



Class.DATE	Frequency	Relative %
2013	58	157
2014	47	116.7
2015	35	85.1
2016	30	73
2005	29	144.3
2011	20	66.3
2010	20	72.6
2008	19	89.7
2009	18	73.4
2017	18	43.2
2018	17	60.7
2012	15	48.7
2006	13	77
2007	13	67.1
0	3	483.1

Rows per page: 500

Class.REGION	Frequency	Relative %
Europe	266	120.3
Asia	51	91.1
North_America	18	24.6
Africa	11	31.8
MENA	5	26.4
Oceania	4	25.2

Class.TYPE	Frequency	Relative %
General_Document	189	216.5
Activity_Report	147	46.3
Strategy	19	101.9

	A	B	C	D
31	class.DATE	"Frequency"	"Relative frequency"	
32	2013	58	157	
33	2014	47	116.7	
34	2015	35	85.1	
35	2016	30	73	
36	2005	29	144.3	
37	2011	20	66.3	
38	2010	20	72.6	
39	2008	19	89.7	
40	2009	18	73.4	
41	2017	18	43.2	
42	2018	17	60.7	
43	2012	15	48.7	
44	2006	13	77	
45	2007	13	67.1	
46	0	3	483.1	
47	class.REGION	"Frequency"	"Relative frequency"	
48	Europe	266	120.3	
49	Asia	51	91.1	
50	North_Amer	18	24.6	
51	Africa	11	31.8	
52	MENA	5	26.4	
53	Oceania	4	25.2	
54	class.TYPE	"Frequency"	"Relative frequency"	
55	General_Doc	189	216.5	
56	Activity_Rep	147	46.3	
57	Strategy	19	101.9	

	A	B	C	D
1	Class	Text Type	Frequency	Relative Frequency
31	Year	2013	58	157
32	Year	2014	47	116.7
33	Year	2015	35	85.1
34	Year	2016	30	73
35	Year	2005	29	144.3
36	Year	2011	20	66.3
37	Year	2010	20	72.6
38	Year	2008	19	89.7
39	Year	2009	18	73.4
40	Year	2017	18	43.2
41	Y	2018	17	60.7
42	Y	2012	15	48.7
43	Year	2006	13	77
44	Year	2007	13	67.1
45	Year	0	3	483.1
46	Region	Europe	266	120.3
47	Region	Asia	51	91.1
48	Region	North_Amer	18	24.6
49	Region	Africa	11	31.8
50	Region	MENA	5	26.4
51	Region	Oceania	4	25.2
52	Doc Type	General_Doc	189	216.5
53	Doc Type	Activity_Rep	147	46.3
54	Doc Type	Strategy	19	101.9

Figure 5: Spreadsheet treatment for a text type report

In Tableau, fields for Class and Text Type are set as dimensions, while absolute frequency and relative frequency are set as measures. As shown in Figure 6, our map represents, for each HE region, frequency with solid colour bubbles and relative frequency with a ring around each bubble. Tableau comes with a great deal of predefined geographical units for disaggregation such as countries, US states, Canadian provinces, European NUTS, among others. However, these do not match HE regions. By means of calculated fields, we linked our HE regions to a country whose location on the map serves as a good anchor point for each bubble. For example, the Europe bubble is anchored to Denmark, whereas the CCSA (Central Caribbean and South America) bubble is anchored to Bolivia.

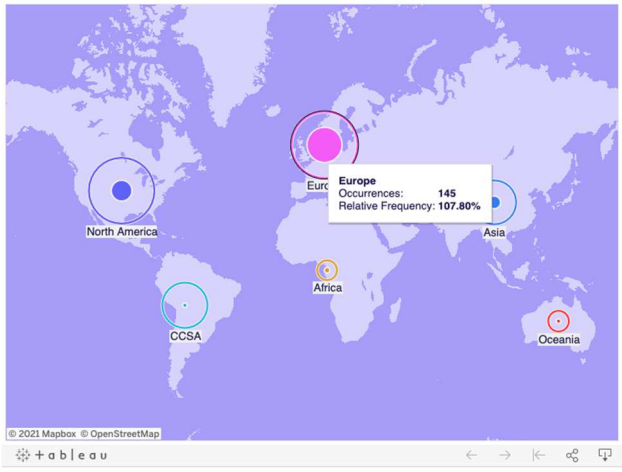


Figure 6: A map for HUMANITARIANISM

The same dataset can be used to build visualisations that compare relative frequency with absolute frequency, disaggregating by text type, namely year, region, organisation and document type. This entails building four different bar charts, which can be presented together with a Tableau story. Figure 7 shows the default view of this story, a histogram representing relative frequency as bars and absolute frequency as a superimposed line. To view the other three disaggregation options, users can use the buttons located at the top.

Exploring the visualisation in Figure 7 in detail sheds light on the temporal evolution of PARTICIPATION. Collectively, its occurrences were highest in 2015, whereas 2013 saw the highest relative frequency with nearly 160 %; European general documents generated the greatest number of occurrences; and the top five organisation types with the highest relative frequency of participation are WHS, C/B, RC, NGO_Fed and Net.

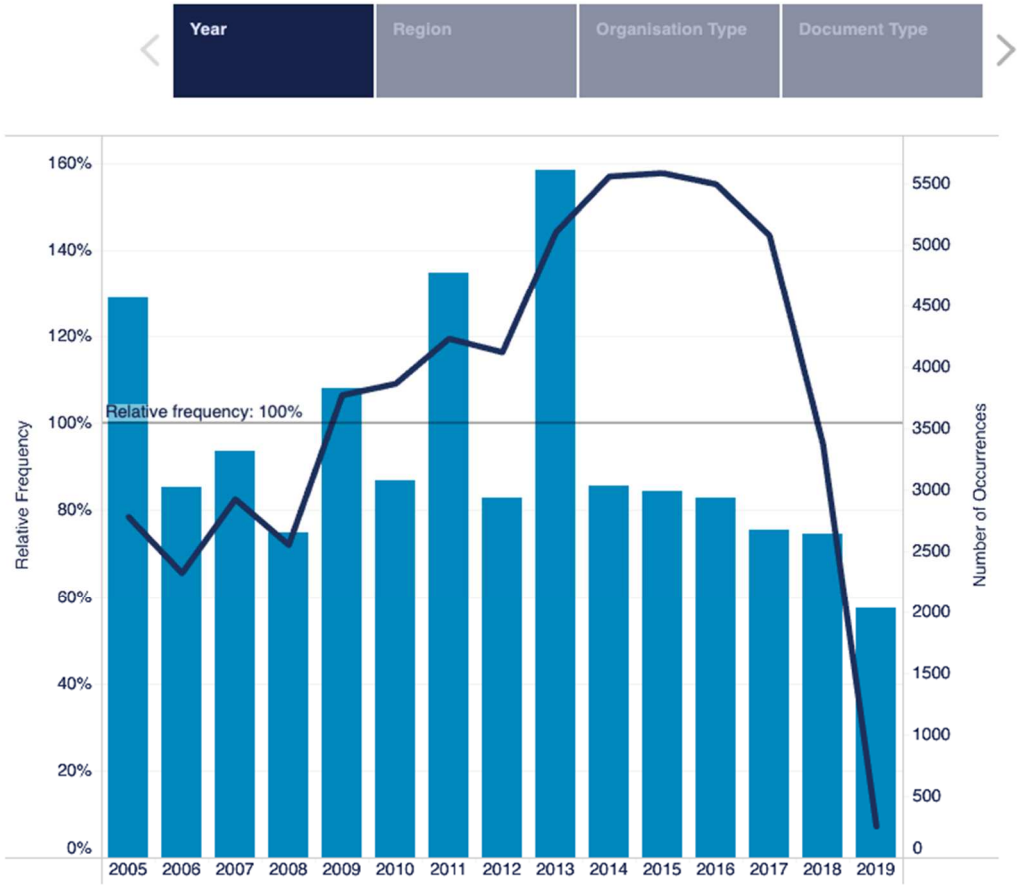


Figure 7: A Tableau story showing a histogram comparing absolute and relative frequency for PARTICIPATION

3.3 Top Yearly Collocate Histogram

To explore the evolution of collocates for a given search expression (which can be highly informative when analysing concept dynamics), a dataset can be built by conducting multiple queries in Sketch Engine. This dataset has to contain the necessary information to disaggregate collocates by year and organisation type.

To begin, we conduct multiple queries so as to obtain collocate reports for each year. This can be achieved by specifying document metadata in each CQL query. Firstly, we query the corpus with the Concordance tool by using the expression [lemma_lc="*x*"] within <class(DATE="*y*")>, where *x* is the term or terms designating a concept, and *y* is a year of publication. Once a list of concordances is generated, we then select the Collocations functionality, which computes collocations of the search term or terms. Even though Sketch Engine encourages the use of its Word Sketch tool for this purpose, there are unfortunately two issues with this. The first is that it can only be used with lemma tags, which means that capitalised occurrences are automatically discarded. The second is that it does not work well with multi-word expressions, which is the case for many of HE concepts. In the Collocations functionality, we set a range of -3, 3 and select lemma (lowercase) as the computation attribute. Finally, a collocational report is generated, which contains all extracted collocates and a set of measures. This step has to be repeated 15 times, changing the year of publication for each query. For the purposes of our dataset, we are interested in the collocates and their corresponding logDice score.

For yearly organisation type-specific collocation reports, we query the corpus with [lemma_lc="*x*"] within <class(DATE="*y*") & (ORGANIZATION TYPE ="*z*" >, where z is the code of the five organisation types with the highest absolute frequency. As with organisation type-unspecific reports, collocational reports are generated through the Collocations functionality with the same settings. This task has to be performed 75 times, with all possible year-organisation type combinations.

Before all individual collocational reports are combined into a single spreadsheet, we curate collocates manually to remove prepositions, truncated words and other empty expressions from the lists. To ease the process, a stop word list is used, which is continuously fed with removed collocates from previous tasks. With spreadsheet software, records from organisation type-unspecific and specific reports are added in the file with an additional column. Furthermore, a second column is added to indicate the year of publication. This leaves us with a single CVS containing collocates by year for the entire corpus, as well as collocates by year disaggregated by organisation type.

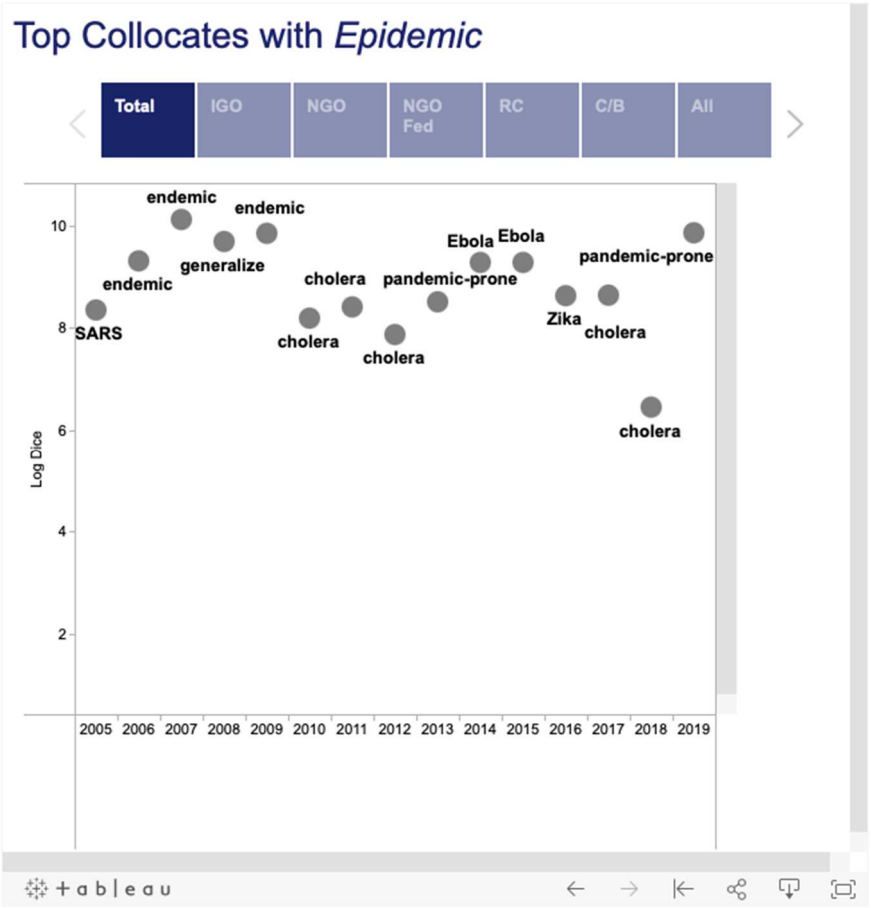


Figure 8: Histogram showing yearly top collocates for EPIDEMIC across the entire corpus

With such a dataset, we can create collocational histograms that allow users to see the top collocate for each year for the entire corpus (i.e., the whole set of concordances), as well as disaggregated by the top five organisation types (i.e., the five organisation types that generate the highest frequencies).

For instance, in the case of EPIDEMIC, epidemic types are the most salient collocates over the years (*SARS* in 2005, *cholera* in 2010-2012, 2017 and 2018, *Ebola* in 2014-15), *zika* in 2016). In 2006-7 and 2009 *endemic* stands out and the single verb in the selection is *generalize* (2008). More recently (2019), *pandemic-prone* is the top collocate, which reflects current concerns about epidemics. *Pandemic-prone* and *pandemic* seem to have been relevant for IGOs and RC for longer (the top collocate in 2013 and 2019 for IGOs' and 2010, 2013 and 2016 for RC). IGO's top collocates related to epidemic types also include *meningitis* (2007, 2009), whereas NGOs show more interest in *malaria* (2007) and *AIDS* (2008, 2009). The only top collocates related to epidemic management are found in texts by NGOs and C/Bs: *combat*, *forecasting* and *prevention*. And the only collocates related to causes are mentioned by NGO_Feds and NGOs: *miningococal* and *waterborne*.

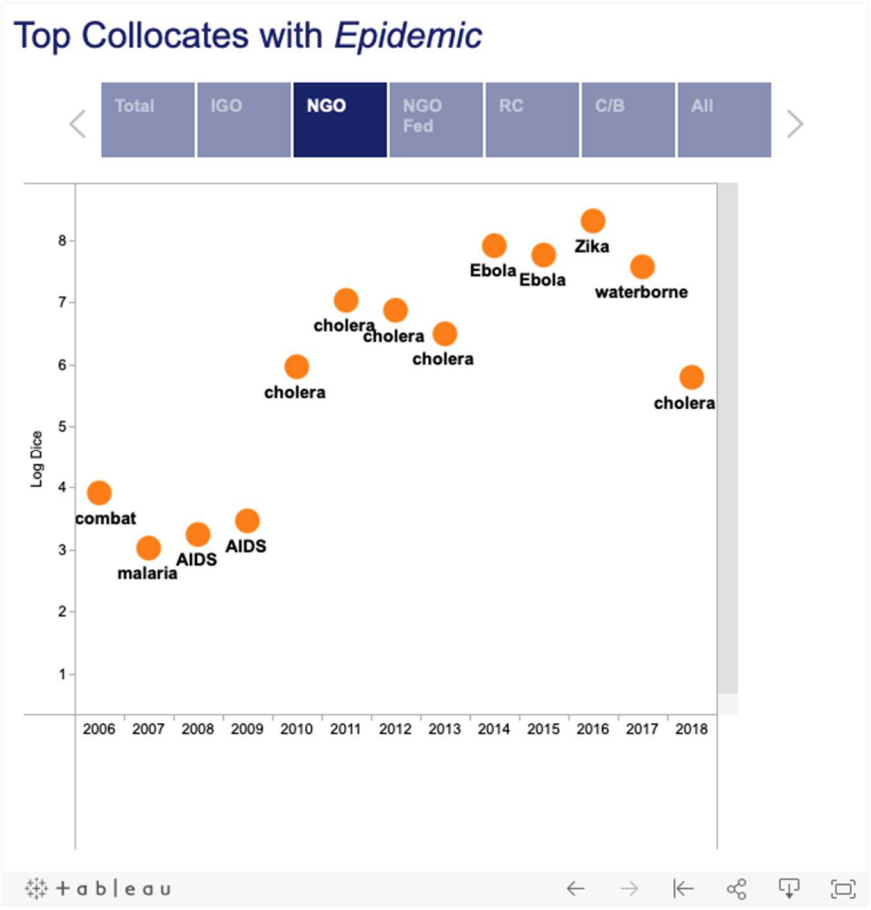


Figure 9: A histogram showing top yearly collocates for EPIDEMIC in NGO documents

The building process in Tableau is similar to that of the relative frequency bar charts (see Section 4.2) in that it requires multiple visualisations be presented together with the Story functionality. Fields for collocates, years and organisation types are set as dimensions, whilst logDice is set as a measure. As can be seen in Figure 8, collocates are presented as colour circles placed in a histogram at varying heights based on their logDice score. Figure 9 shows the top yearly collocates obtained from NGO documents.

3.4 Unique and Shared Collocates

Reporting on collocates that are unique to a single organisation type constitutes a way of ascertaining what a given organisation says about a concept that others do not. Examining which collocates are shared by multiple organisations can help identify what common areas among organisations when discussing a certain concept.

A dataset for this purpose can be built by corpus querying with a similar method seen in Section 4.3. However, in this case, we use the CQL expression

[lemma_lc="*x*"]within<class(ORGANIZATION_TYPE="*y*")>, which does not specify a year of publication. The rest of the extraction process in Sketch Engine is identical. The corpus has to be queried five times for each of the five organisation types with the highest frequencies.

To combine the five collocational reports into CSV, a column is added in the spreadsheet to specify the organisation type from which each record was obtained. After this process, we have a dataset that enables us to compare collocates among organisation types. In the same workbook in Tableau, the fields for collocates and organisation types are set as dimensions, whereas logDice is set as a measure. Collocates unique to each organisation type can be well represented with a packed bubble chart (Figure 10). By means of a conditional set, collocates found in more than one organisation type can be filtered out.

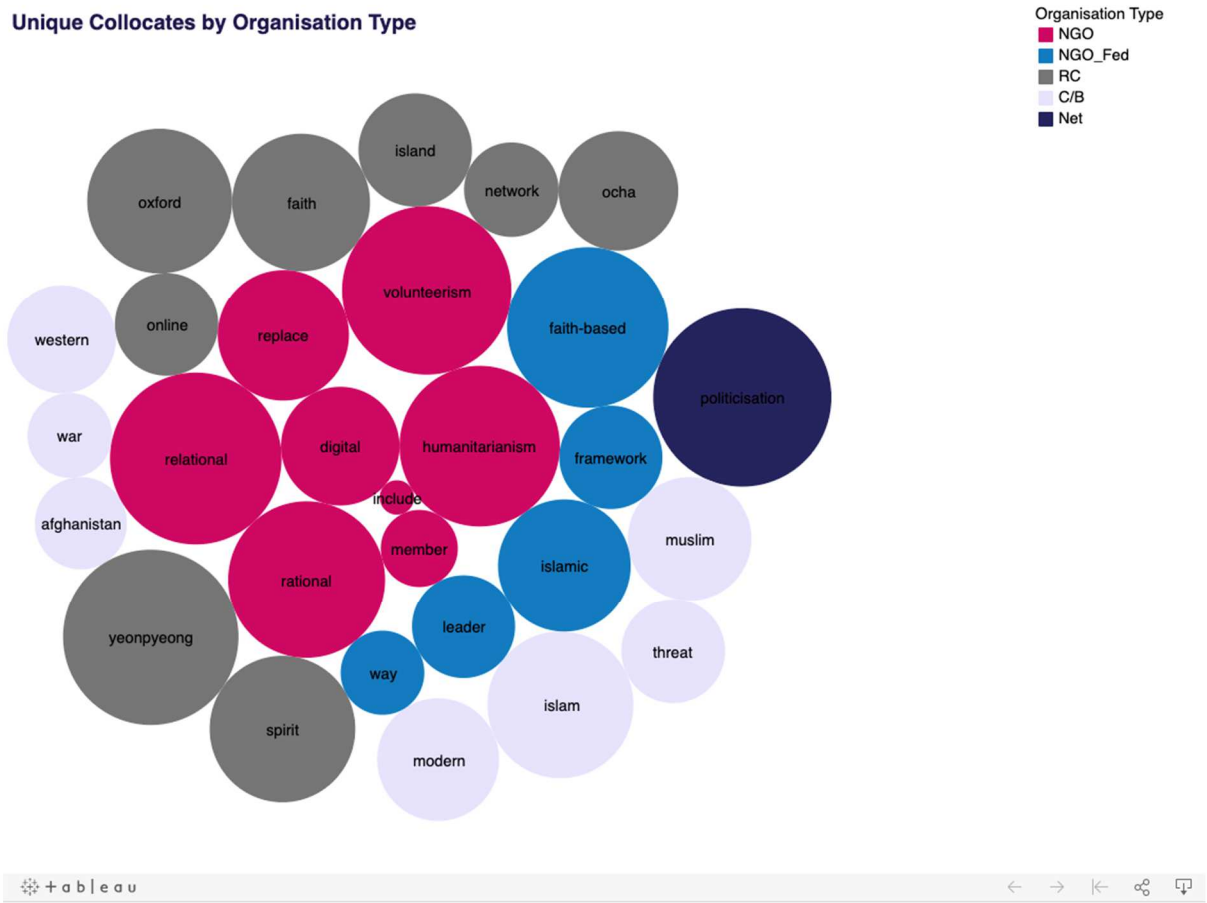


Figure 10: Unique collocates for HUMANITARIANISM

Collocates shared by multiple organisation types would be optimally represented by Venn diagrams. Here, shared collocates can be understood as the collocates that constitute intersections between organisation types, i.e., intersections between subcorpora. However, Tableau does not offer an option to build Venn diagrams. For this reason, we resorted to bar charts, which serve as a good alternative. With a parameter, a dynamic conditional set and a filter, we can create a bar chart that

shows which collocates are shared by two or more organisation types. As can be seen in Figure 11, each collocate is represented by a bar that can be divided into multiple colour sections. The colour of each section represents an organisation type, while its size represents the collocate's logDice score within that given type. Thanks to a filtering parameter, users can filter collocates by the number of organisation types in which they appear.

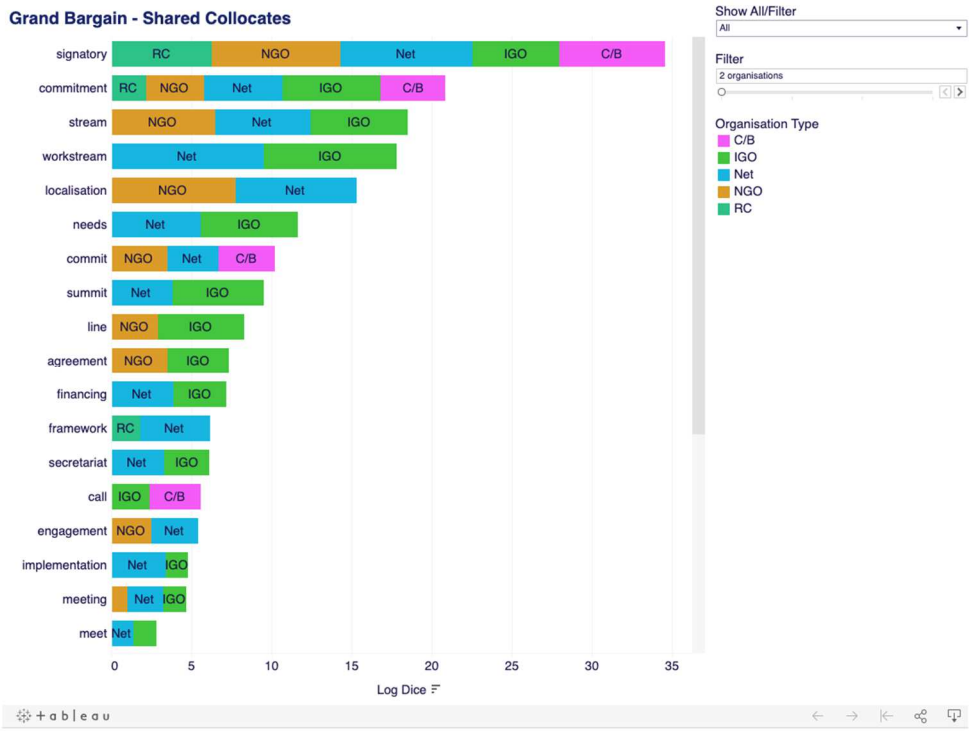


Figure 11: Shared collocates for GRAND BARGAIN

Thanks to the visualisations in Figure 10 and 11, linguists can inform experts about the collocating trends of *humanitarianism*. For instance, NGO documents feature the unique collocates of *relational*, *volunteerism*, *rational*, *replace*, *digital*, *member* and *include*, whereas C/B documents contain *Islam*, *Muslim*, *modern*, *Western*, *threat*, *Afghanistan*, *war* and *project*, pointing to very different concerns. The top collocates shared by two organisations are *value* and *crisis*, whereas the only collocate shared by three organisations is *development*. No collocates were found to be shared by either four or five organisations, which could indicate that large discrepancies are found in the conceptualisation of HUMANITARIANISM.

3.5 Square Treemaps

Square treemaps are an interesting option for the categorisation of multiple elements, as well as measures associated with said elements. This section will examine two case uses of this visualisation option within Tableau.

3.5.1 Representing compound concepts

Complex nominals are phrases consisting of a head noun modified by other elements, such as other nouns, adjectives and prepositional phrases. They are considered as instantiations of conceptual combinations, whereby compound concepts are formed by pre-existing simpler ones (Cabezas-García & Chambó, in press). Analysing the understanding of a concept in a given domain requires looking at the conceptual combination that it forms. Square treemaps are an effective way of representing such information.

MWterms_modifier			
faith_leader faith leaders	267	11.81	•••
faith_community faith communities	173	11.24	•••
faith_group faith groups	131	10.87	•••
faith_actor faith actors	56	9.69	•••
faith_based_organization faith based organizations	32	8.87	•••
faith_based_organisation faith based organisations	26	8.54	•••
faith_tradition faith traditions	21	8.3	•••
faith_perspective from a faith perspective	19	8.16	•••
faith_organisation faith organisations	13	7.61	•••
faith_formation faith formation	13	7.61	•••
local_faith_community local faith communities	12	7.49	•••
faith_network faith networks and	11	7.37	•••
people_of_different_faiths among people of different faiths	11	7.37	•••
other_faith_group with other faith groups	11	7.37	•••
other_faith_community with other faith communities	11	7.37	•••

Figure 11: Word Sketch for MWTs for FAITH

Given that FAITH is designated by a monolexical term, we used a modified version of Sketch Engine's default sketch grammar, which is the backbone of the Word Sketch tool. This custom sketch grammar is able to extract the multi-word terms (MWTs) in which the search term appears as both as a head or a modifier. On the one hand, MWTs with *faith* as a head constitute hyponyms of FAITH (e.g., CHRISTIAN FAITH, ISLAMIC FAITH, LOCAL FAITH, etc.), which can also be classified

according to different facets. On the other hand, MWTs with *faith* as a modifier constitute conceptual combinations in which faith intervenes (e.g., FAITH LEADER, FAITH COMMUNITY, FAITH IDENTITY, etc.), which would point to non-hierarchical relations and event participants. To represent the conceptual compounds with faith contained in the HE corpus, we extracted the MWTs with *faith* as a modifier (Figure 12).

All extracted MWTs with their frequencies were transferred into a spreadsheet and classified into conceptual categories by creating additional columns. Separately, another spreadsheet was manually populated with sample contexts from the HE corpus for each MWT, together with each context's metadata.

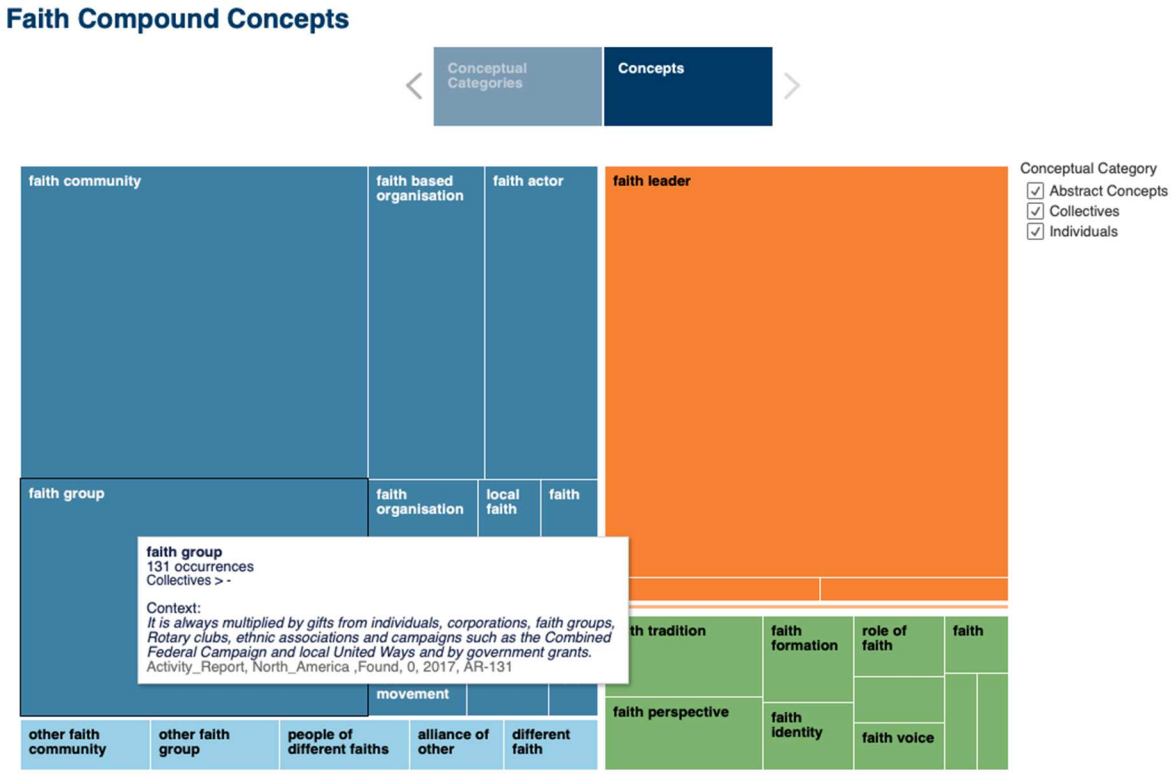


Figure 12: Square treemap providing
a summary of compound concepts with FAITH

In Tableau, both spreadsheets are joined with a union. The frequency for each MWT is set as the measure, while compound concepts and context metadata are set as dimensions. In a square treemap, each compound concept is symbolised by a rectangle whose size represents its frequency in the corpus. As can be seen in Figure 12, when the user hovers a rectangle, a tooltip provides a sample context as well as the details of the document from which it was sourced.

3.5.2 Representing coordinated concepts

In text, associated concepts may also appear in coordination. The Word Sketch functionality can extract expressions linked to a search term through coordination

with the conjunctions *and* and *or*. However, it is not powerful enough to extract coordinated MWTs. For this reason, in order to create a dataset containing all coordinated concepts with HUMANITARIANISM, we queried the corpus with the following two CQL expressions:

– [tag="N.*|J.*"]{1,3} within ([lemma_lc="humanitarianism"] [word="and|/or"] ([tag="N.*|J.*"]{1,3}within[tag!="N.*|J.*"][tag="N.*|J.*"]{1,3}[tag!="N.*"]))

– [tag="N.*|J.*"]{1,3} within (([tag="N.*|J.*"]{1,3}within[tag!="N.*|J.*"][tag="N.*|J.*"]{1,3}[tag!="N.*"]) [word="and|/or"] [lemma_lc="humanitarianism"])

In brief, the above expressions extract both single-word and multiword expressions coordinated with humanitarianism. Concordances were filtered with the Hide Sub-Hits quick filtering functionality, which removes concordances including partial hits. This is bound to occur when using ranges (e.g., {1,3}) to capture complex nominals. Both sets of concordances were computed using the Frequency functionality, which generated two report containing full coordinated expressions on both sides of our search term.

Figure 12: Coordinated concepts with HUMANITARIANISM

Both reports were combined into a single spreadsheet containing frequencies for each expression. As with the case use described in Section 4.5.1, a separate spreadsheet with context samples was also built. Similarly, both data sources were joined with a union in Tableau and visualised using the treemap functionality. The

resulting visualisation provides a summary of the concepts coordinated with HUMANITARIANISM (Figure 12). The analysis of conceptual compounds reveals that humanitarian discourse is concerned with notional discussions about the concept of HUMANITARIANISM. Other important aspects include the *constituent elements* of humanitarianism (e.g. core values, language, activities, practice, etc.) and those *processes that affect humanitarianism* (e.g. demilitarisation, politicisation, sanctification, etc.).

3.6 Conceptual Development Histogram

Some concepts can be so specific that it pays to represent their development over time. This is usually the case for compound concepts that generate a handful of knowledge-rich contexts, which can be curated manually and classified into descriptive categories. The compound concept of ACCOUNTABILITY TO AFFECTED POPULATIONS is a highly specialised humanitarian concept that is formed by AFFECTED POPULATION, a concept with constitutes a fully-fledged entry in the HE. Numerous occurrences of its acronym – AAP – Indicate that the concept has solidified.

A low number of occurrences allows a linguist to download and classify statements manually into multiple categories, thus creating a heavily textual dataset. Using a similar method as described in Section 4.1., a histogram can be built to represent categorised contexts by year as shown in Figure 13.

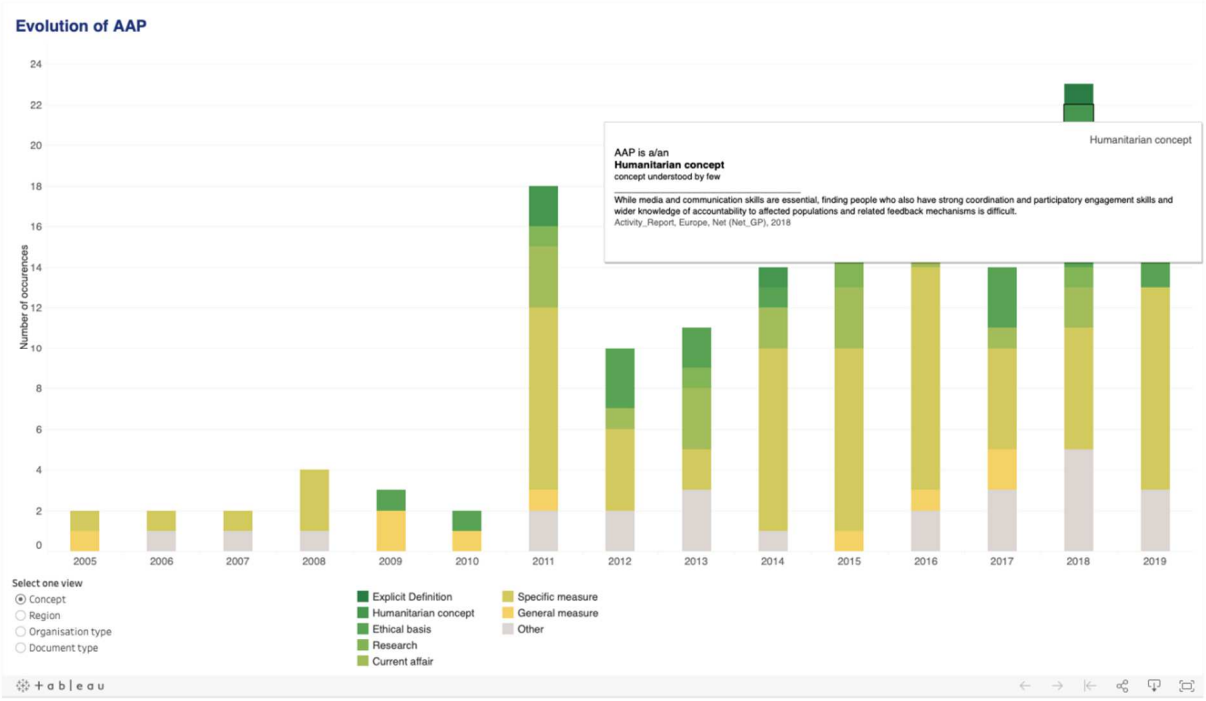


Figure 13: A histogram representing the evolution of ACCOUNTABILITY TO AFFECTED POPULATIONS (AAP)

All contexts were classified into eight statement categories based on what organisations say about it, namely: general measure, specific measure, current affair, research, ethical basis, humanitarian concept, explicit definition and other. Contexts in the general measure and specific measure categories describe measures taken or that could be taken by organisations to increase ACCOUNTABILITY TO AFFECTED POPULATIONS. Contexts categorised as a current affair describe the concept as an ongoing concern in the humanitarian domain. The research category includes contexts stating that research is either needed or being conducted to increase and/or better understand ACCOUNTABILITY TO AFFECTED POPULATIONS. The ethical basis category consists of contexts in which the concept is described as an organisational value or principle. Contexts in the humanitarian concept category state that it constitutes a humanitarian concept, whilst the explicit definition category contain an authoritative definition for the concept as found in the corpus. Lastly, contexts classified as other mostly include statements merely claiming that a given organisation works towards AAP, as well as other marginal cases. This statement classification system makes it possible to represemt how AAP develops from vague mentions to more specific and defined mentions. In the visualisation, hovering over each bar section to reveals more details about each mention, including the context.

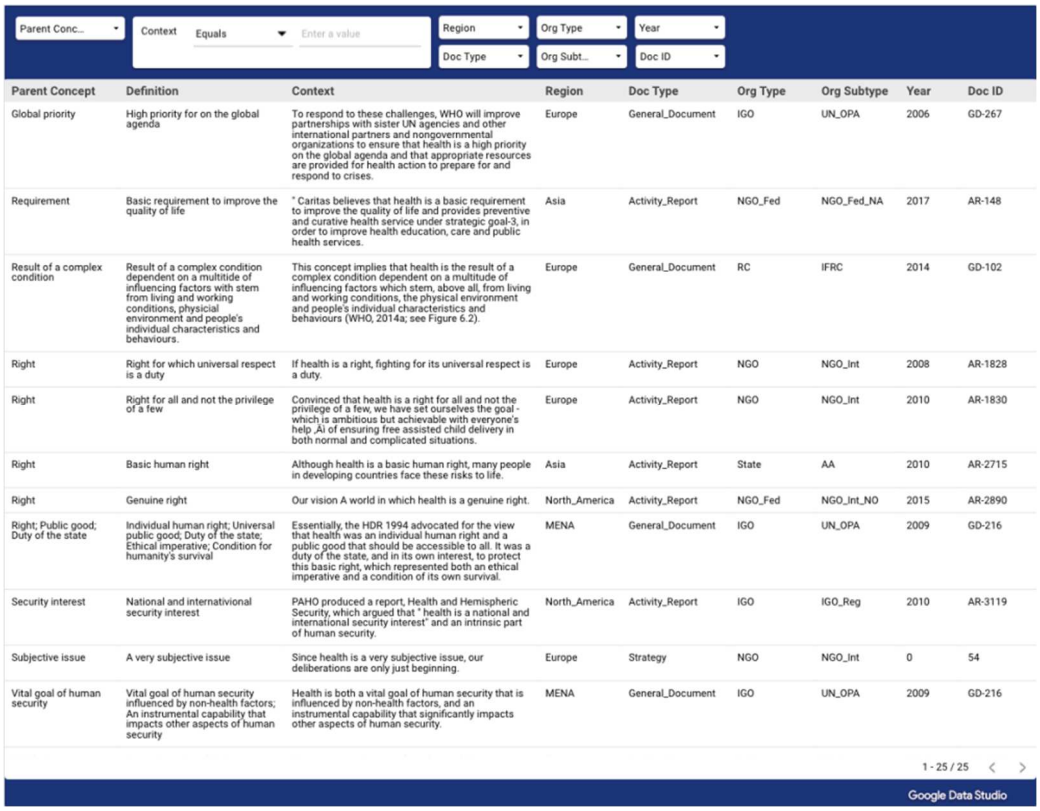
In 2011, APP began to attract attention as evidenced by a dramatic increase in occurrences. Documents published in this year contain the two first mentions pointing out that AAP is an ill-defined concept. It was not until 2015 that a progressive increase surpasses the value for 2011. Mentions of conceptual vagueness reappeared in 2015 and 2016. The greatest number of occurrences were obtained from documents published in 2018, when a change in proportion of statement categories can be observed. 2018 saw the only explicit definition of AAP as well as a considerable increase of occurrences classified as ethical basis. 2019 also experiences a change in proportion, with most statements being from the specific measure and ethical basis categories. It also ceased to be referred to as a cross-cutting issue or key current challenge.

Analysis suggests that ACCOUNTABILITY TO AFFECTED POPULATIONS crystallised as a concept in 2018. This is when its first definition appeared, and the greatest number of organisations claimed their adherence to it as a principle.

3.7 Filterable and Searchable Tables

Sometimes linguistic reporting for certain concepts may require presenting entire sets of manually curated contexts. Reporting on explicit definitions is perhaps the first step when describing the conceptualisation of a notion. For example, the HE corpus contains many definitions for the concept of HEALTH in varying degrees of

explicitness. As shown in Figure 14, these were presented in a sortable and searchable table built with Google Data Studio. These are designed to allow users to search and filter contexts.



Parent Concept	Definition	Context	Region	Doc Type	Org Type	Org Subtype	Year	Doc ID
Global priority	High priority for on the global agenda	To respond to these challenges, WHO will improve partnerships with sister UN agencies and other international partners and nongovernmental organizations to ensure that health is a high priority on the global agenda and that appropriate resources are provided for health action to prepare for and respond to crises.	Europe	General_Document	IGO	UN_OPA	2006	GD-267
Requirement	Basic requirement to improve the quality of life	" Caritas believes that health is a basic requirement to improve the quality of life and provides preventive and curative health service under strategic goal-3, in order to improve health education, care and public health services.	Asia	Activity_Report	NGO_Fed	NGO_Fed_NA	2017	AR-148
Result of a complex condition	Result of a complex condition dependent on a multitide of influencing factors with stem from living and working conditions, physicial environment and people's individual characteristics and behaviours.	This concept implies that health is the result of a complex condition dependent on a multitude of influencing factors which stem, above all, from living and working conditions, the physical environment and people's individual characteristics and behaviours (WHO, 2014a; see Figure 6.2).	Europe	General_Document	RC	IFRC	2014	GD-102
Right	Right for which universal respect is a duty	If health is a right, fighting for its universal respect is a duty.	Europe	Activity_Report	NGO	NGO_Int	2008	AR-1828
Right	Right for all and not the privilege of a few	Convinced that health is a right for all and not the privilege of a few, we have set ourselves the goal - which is ambitious but achievable with everyone's help ¸Äi of ensuring free assisted child delivery in both normal and complicated situations.	Europe	Activity_Report	NGO	NGO_Int	2010	AR-1830
Right	Basic human right	Although health is a basic human right, many people in developing countries face these risks to life.	Asia	Activity_Report	State	AA	2010	AR-2715
Right	Genuine right	Our vision A world in which health is a genuine right.	North_America	Activity_Report	NGO_Fed	NGO_Int_NO	2015	AR-2890
Right; Public good; Duty of the state	Individual human right; Universal public good; Duty of the state; Ethical imperative; Condition for humanity's survival	Essentially, the HDR 1994 advocated for the view that health was an individual human right and a public good that should be accessible to all. It was a duty of the state, and in its own interest, to protect this basic right, which represented both an ethical imperative and a condition of its own survival.	MENA	General_Document	IGO	UN_OPA	2009	GD-216
Security interest	National and internativional security interest	PAHO produced a report, Health and Hemispheric Security, which argued that " health is a national and international security interest" and an intrinsic part of human security.	North_America	Activity_Report	IGO	IGO_Reg	2010	AR-3119
Subjective issue	A very subjective issue	Since health is a very subjective issue, our deliberations are only just beginning.	Europe	Strategy	NGO	NGO_Int	0	54
Vital goal of human security	Vital goal of human security influenced by non-health factors; An instrumental capability that impacts other aspects of human security	Health is both a vital goal of human security that is influenced by non-health factors, and an instrumental capability that significantly impacts other aspects of human security.	MENA	General_Document	IGO	UN_OPA	2009	GD-216

Figure 12: A filterable and searchable table showing a selection of explicit definitions for HEALTH

These tables are presented in a separate subpage and are mainly intended as supportive evidence for a linguist's claims in the many pages and body of his or her LAR. By analysing explicit definitions, the linguist concludes that definitions are built on three distinct conceptualisations of HEALTH: health as a state or condition, health as human right; and health as a fundamental component.

These tables are also used in order to store and provide an interactive access to debates and controversies, widely found in humanitarian discourse and manually curated and categorised by HE linguists.

3.8 Recapitulation

Table 5 provides a summary of all the visualisation types discussed in this paper. It also contains a link to each visualisation on Tableau public and Google Data Studio from which it can be freely downloaded.

Visualisation	**Purpose**	**Dimensions**	**Measures**
Frequency Histogram	To display the evolution of frequency over time, allowing users to disaggregate yearly frequencies by organisation type, document type and region.	Year, Organisation type, Region, Document type	Frequency
Map	To display the geographical distribution of absolute frequency and relative frequency among regions.	Region	Frequency, Relative frequency
Relative Frequency Bar Chart	To compare yearly absolute frequencies and relative frequencies, allowing users to explore other distributions by organisation type, document type and region.	Year, Organisation type, Region, Document type	Frequency, Relative frequency
Top Yearly Collocate Histogram	To compare most significant collocates over time and among organisation types.	Lexical unit (collocate), Year, Organisation type	logDice
Unique Collocates	To show collocates unique to organisation types.	Lexical unit (collocate), Organisation type	logDice
Shared Collocates	To show collocates shared by two or more organisation types.	Lexical unit (collocate), Organisation type	logDice
Compound Concept Treemap	To provide a summary of the lexical compounds in which a search expression intervenes, arranged by semantic categorisation and frequency.	Lexical unit (compound), Category, Context, Year, Organisation type, Region, Document type	Frequency
Coordinated Concept Treemap	To provide a summary of the lexical units appearing in coordination with a search expression, arranged by frequency.	Lexical unit, Context, Year, Organisation type, Region, Document type	Frequency
Conceptual Evolution	To display the evolution of the conceptualisation of a notion by	Context, Hypernym, Context, Category, Year,	None

Histogram	arranging and categorising contexts with varying degrees of definitional precision.	Organisation Type, Region, Document Type	
Filterable and Searchable Table	To display a set of manually curated contexts.	Hypernym, Definitional element, Context, Region, Organisation type, Organisation subtype, Document type, Document ID	None

Table 5: Summary of visualisations

4. Conclusions and future work

In this paper we have shown how data visualisation can have a two-fold role in a corpus-driven project. It can assist linguists for the interpretation of corpus information in a field where they are not experts, but it can also be especially useful when serving as intermediary with field experts. Field experts, who are not familiar with corpus linguistics or raw lexical data, can benefit from interactive visualisations because they can freely interact with the data in a more intuitive fashion and build their own claims, complementing those offered by the team of linguists.

Most entries in the HE are expected to be written by external experts. Nonetheless, linguist-expert interaction is still limited to an in-house humanitarian at the HE. At the time of writing, only one LAR has been used to build a sample entry, which served to validate the LAR-building process. This will also provide external experts with a reference for guidance when writing their own entries. In addition, linguists are also interacting with another in-house expert who is in charge of compiling a list of concept-specific research questions. Sometimes, these questions may be answered by querying the corpus. This form of linguist-expert interaction provides linguists with concept-specific tasks and therefore contributes to shaping each LAR by adding particularised sections. As content production is expected to scale up, we will soon have more data on linguist-expert interaction, which will prompt a new line of research and provide us with a new way to improve our data visualisation skills.

In parallel, our efforts are currently centred on designing visualisations that represent collocational intersections between subcorpora more satisfactorily. For example, Venn diagrams with RStudio have the potential to replace our current packed bubble charts in future LARs. Additionally, we are working on a system to query the HE Corpus through Sketch Engine's API. At present, collocational data is only being extracted from the top five organisation types with the greatest

number of occurrences of a given term. To create histograms, collocates are also disaggregated by year of publication. More meaningful comparisons between subcorpora could be drawn if collocational data were further disaggregated by every type of corpus metadata, i.e. increasing granularity. With our current manual approach, our top yearly collocate histogram for one concept requires a total of 90 queries through Sketch Engine's graphic user interface. Using Sketch Engine's API will not only remove manual querying tasks from our workflow, but it will also provide us with richer and more comprehensive datasets.

5. Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.

6. References

Allen, W. (2017). Making corpus data visible: Visualising text with research intermediaries. Corpora, 12(3), pp. 459–482. Available at: https://doi.org/10.3366/cor.2017.0128

Cabezas-García, M. & Chambó, S. (in press). Multi-Word Term Variation: Prepositional and Adjectival Complex Nominals in Spanish. SJAL (Spanish Journal of Applied Linguistics).

Christ, O., Schulze, B.M., Hofmann, A., & König, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP) - User's Manual. Institute for Natural Language Processing. Stuttgart: University of Stuttgart.

Guillaume D. (2019), Mapping lexical variation with Tableau software. Around the word. Accessed at: https://corling.hypotheses.org/2853 (25 March 2021)

Humanitarian Encyclopedia (2020). Available at: https://humanitarianencyclopedia.org/wp-content/uploads/2020/05/HE-brochure_finalMay2020.pdf

Kilgarriff, A. Rychlý P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) Proceedings of the 11th EURALEX International Congress. EURALEX. Lorient, France, pp 105–115.

León-Araúz, P. & San Martín, A. (2018) The EcoLexicon Semantic Sketch Grammar: from Knowledge Patterns to Word Sketches. In I. Kernerman & S. Krek (eds.) Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets. Miyazaki, Japan, pp. 94–99. Available at: http://lexicon.ugr.es/pdf/Leon-Arauz2018.pdf

Linguistic Analysis Portal for the Humanitarian Encyclopedia Available at: https://sites.google.com/view/humanitarianencyclopedia

Rychlý P. (2008). A lexicographer-friendly association score. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6–9

Siirtola, H., Räihä, K. J., Säily, T., & Nevalainen, T. (2010). Information visualization for corpus linguistics: Towards interactive tools. International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 33–36. Available at: https://doi.org/10.1145/2002353.2002365