A recommendation to the SSH community: Take a linguist on board

Jeannine Beeken[1]

Abstract

In this paper we address how Natural Language Processing (NLP) approaches and language technology can contribute to data services in different ways; from providing social science users with new approaches and tools to explore oral and textual data, to enhancing the search, findability and retrieval of data sources. By using linguistic approaches we are able to process data, for example using Automated Speech Recognition (ASR) and named entity recognizers (NER), extract key concepts and terms, and improve search strategies. We provide examples of how computational linguistics contribute to and facilitate the mining and analysis of oral or textual material, for example (transcribed) interviews or oral histories, and show how free open source (OS) tools can be used very easily to gain a quick overview of the key features of text, which can be further exploited as useful metadata.

Keywords

Natural Language Processing (NLP), language technology, data and metadata services and infrastructure, Social Sciences and Humanities (SSH)

Introduction

Introducing (more) linguistics into social science research is no simple feat. Moreover, the fact that both oral and written material has to be considered presents another complicating factor.

In the following sections, we investigate how (computational) linguistics and tools can contribute to social science research by providing new approaches and tools and by enhancing findability and retrieval, taking into account the fact that these principles promote machine-actionability. We will especially pay attention to findability (online searchable and discoverable) and interoperability (using for example standards and schema, controlled vocabularies, keywords, thesauri or ontologies) of metadata and data.

A basic suite of linguistic tools and methods

In this section, we introduce a possible suite of different linguistic tools and methods which could, or rather, should be added to the standard package of services (for external users) and infrastructure (internal metadata and data managers) at archives.

1. NLP tools which optimize search, findability and retrieval, are spell checkers and correctors, stopword excluders, autosuggest functionality based on one or more thesauri, clustering of keywords and their synonyms (for example 'war' and 'armed conflict'), priority lists of abbreviations/acronyms as used in social science research (for example 'ALS, EHS, CLOSER, GUS') and language-specific stemmers (searching for 'tax' finds correctly studies about 'tax, taxes, taxation', but not about 'taxi' or 'taxis').
2. Automated speech recognition (ASR) tools can partially take over manual transcription, while separating oral language from, for example, silences. They are able to distinguish spoken natural language from surrounding noise and can also recognize and distinguish

between different speakers taking part in, for example, an interview. Most importantly, they convert spoken language into written language or text, using word segmentation, whereby the written text has been aligned with the spoken fragments.

3. Named entity recognizers (NER) identify and classify named entities, such as person names, organisations, geospatial terminology etc. They can simplify anonymization and assist checks on disclosure and de-identification. This feature could be introduced as part of, for example, a self-depositing portal or an ingesting infrastructure.
4. Information extraction (IE) tools detect keywords, thus assisting indexing and pre-populating the relevant metadata fields. Also, this feature could be useful when self-depositing or when adding keywords during the ingesting process. Moreover, matching or aligning the extracted keywords with thesaurus terms or any other (standardized) controlled vocabulary of topics, for example, would improve findability and retrieval.
5. Concordances (KWiC/Keywords in Context) and correlations in a (group of) texts can easily be generated. This feature helps detect possible and unexpected clusters and patterns, for example between 'schools' and 'knives', which leads to new research questions and insights.

Basic linguistic challenges

When developing tools and services for finding and retrieving archived oral and written data, linguists face three main challenges:

1. Disambiguation or separation of, for example, homonyms such as 'book' (as in 'book a holiday' or 'reading a book'), 'bank' (as in 'a bank of a river', 'a savings bank', 'a bank of snow'), 'current' (as in 'my current job', 'ocean currents', 'a current flowing to a lamp')
2. Clustering or grouping of, for example, synonyms (such as 'current, contemporary, present-day, present, ongoing'), words sharing the same stem (such as 'nurse, nurses, nursing'), co-references (such as 'I voted for Sammy, since she is my sister and the current chairperson of the board'), multiword expressions and idioms (such as 'black money', 'black widow', 'black humour', 'with respect to', 'giving the cold shoulder', 'being all ears').
3. Reduction or control, for example removing 'meaningless' stopwords such as 'was, be, you, me, to, for, or, if, when' (excluding 'not, n't') while keeping 'meaningful' keywords and domain dependent or domain specific terminology, for example 'was' or 'WAS' meaning 'Wealth and Assets Survey'.

As said, introducing more linguistics into social science research is no simple feat. For example, some of the challenges mentioned appear to contradict each other—separation, but at the same time also grouping— and the fact that both oral and written data have to be taken into account presents another complicating factor. For example, how do ASR tools (speech to text) distinguish between homophones such as 'be' and 'bee', 'meet' and 'meat', 'friar' and 'fryer', 'nun' and 'none', 'grease' and 'Greece', 'knead' and 'kneed' or 'need'.

Improving search, findability and interoperability by using language-specific tools

Search, findability and retrieval can be optimized by implementing well-known NLP tools such as spell checkers or spell correctors, taking into account case- and diacritic-insensitivity. Other tools that are commonly used within the wide domain of Search Engine Optimization (SEO) provide autosuggestions based on a (domain specific) thesaurus or a controlled list of keywords and their synonyms, which improves findability to a great extent. For example, in the UK Data Service Data

Catalogue, searching for 'health' generates the autosuggested list 'health and well-being', 'health behaviour', 'health care' and a list of thesaurus (in this case HASSET) keywords 'public health risks', 'men's health'. These keywords form an opening to clusters of (quasi-)synonyms (i.e., different form, similar meaning), for example, 'war' and 'armed conflict', or 'energy prices' and 'energy tariffs', 'fuel prices'.

Next, an extended list of acronyms (for example 'CLOSER') and abbreviations (for example 'EHS') as used in social science research can be implemented as part of the search algorithm, giving a higher scoring priority to domain-specific terminology and conceptual meaning than to common language meaning. For example, 'GUS', 'CLOSER', 'DOTS', 'ALS' or 'EHS have a different meaning in social science research (UK Data Service Data catalogue) than in everyday language. Meant is 'Growing up in Scotland'; 'Cohort and Longitudinal Studies Enhancement Resources'; 'IMF Direction of Trade Statistics'; 'Active Lives Survey';' English Housing Survey'; but not 'Global University Systems' or a person's name; 'more close'; a type of punctuation marks (namely '...'); 'Amyotrophic Lateral Sclerosis'; 'Environment, Health and Safety'. When searching for well-known abbreviations and acronyms of datasets/studies and series, for example 'QLFS', 'CLOSER', 'GBHD' or 'WAS', the search engine searches for both the acronym/abbreviation and the full term of the study or dataset/series in the relevant metadata fields. Figure 1. shows the results for studies, when searching for 'closer', sorted by 'Relevance'.

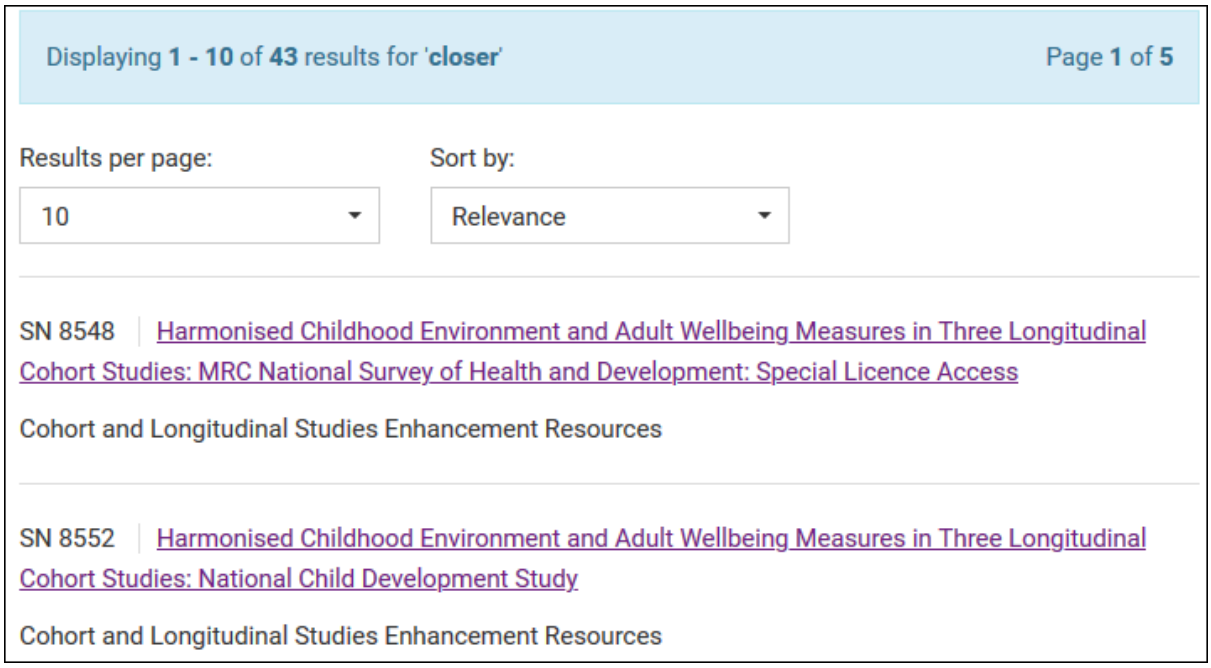

Figure 1.

Moreover, implementing a standardized list of stopwords as an extendable blacklist prevents the retrieval of too many and unwanted results. It therefore improves recall/precision to a high extent. Common stopwords are, for example, 'a, the, by, your, was, she, be, here'. An average list contains between 150 and 250 stopwords. An example: searching the UKDS Data Catalogue for 'was' only retrieves results connected to the 'Wealth and Assets Survey,' and excludes all instances of 'was' as

used in, for example, 'it was'. Consequently, the system retrieves correctly 21 studies in UKDS Data Catalogue (see Figure 2.), whereas without using a list of stopwords, the number of results would be 5000+ due to the appearance of the verb 'was' in for example the abstracts of studies.

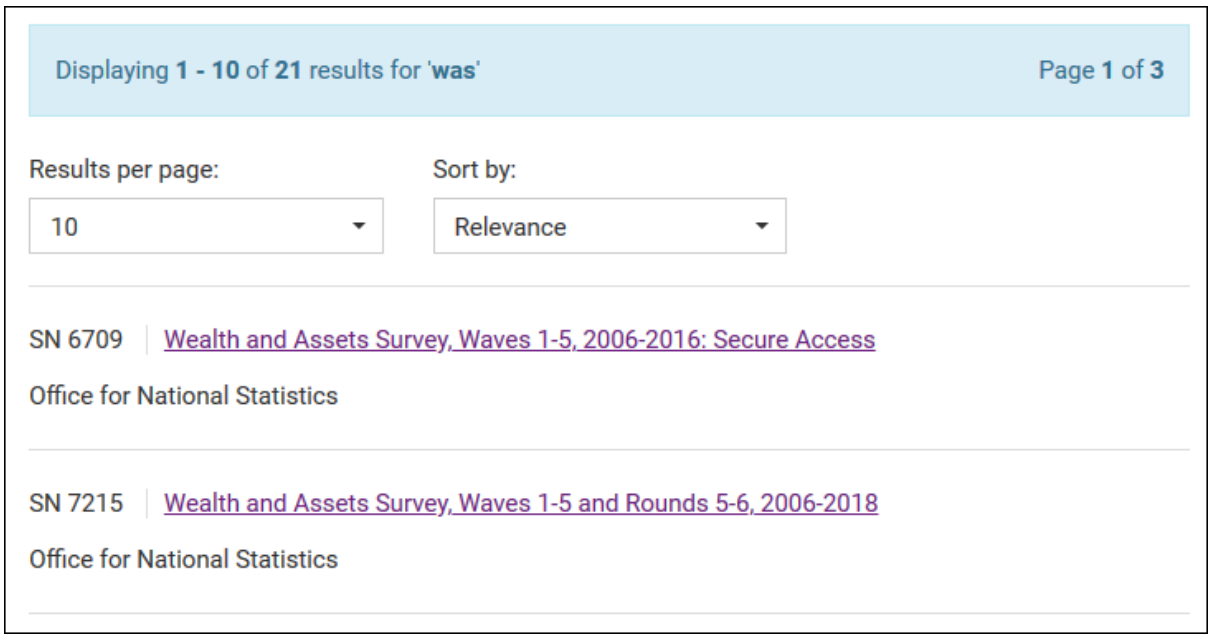

Figure 2.

Findability and retrieval are also improved by implementing language-specific stemmers and lemmatizers. These tools automatically search for all formal (form) and semantic (meaning) variants of a search term, i.e. all the terms that share the same stem or are variants of the canonical form as found in a dictionary. For example, a search for 'nurse' equals a search for 'nurse, nurses, nursing', but does not consider or retrieve studies for 'nurseries'. Lemmatizers help with searches for 'good, better, best', where not all variants share the same stem. Consequently, searches for singular or plural terms and for gerunds yield the same number of results, for example, 'tax' or 'taxes'; 'nurse', 'nurses' *and* 'nursing'. The sorting order of the results, however, will correspond with the specific search term. When searching for 'tax', the results containing the singular form will be higher up in the ranking than the results with the plural form. When searching for 'taxes', the results containing the plural form will appear first (see figures 3. And 4.).

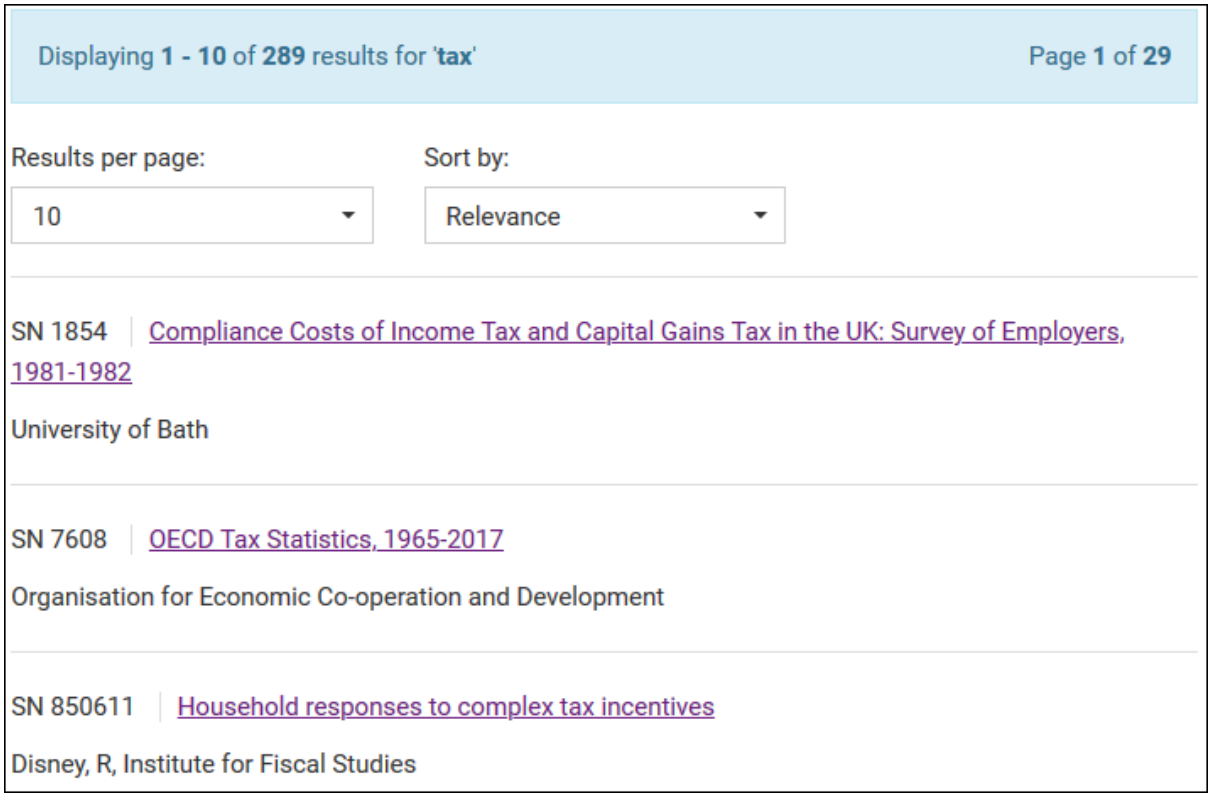

Figure 3.

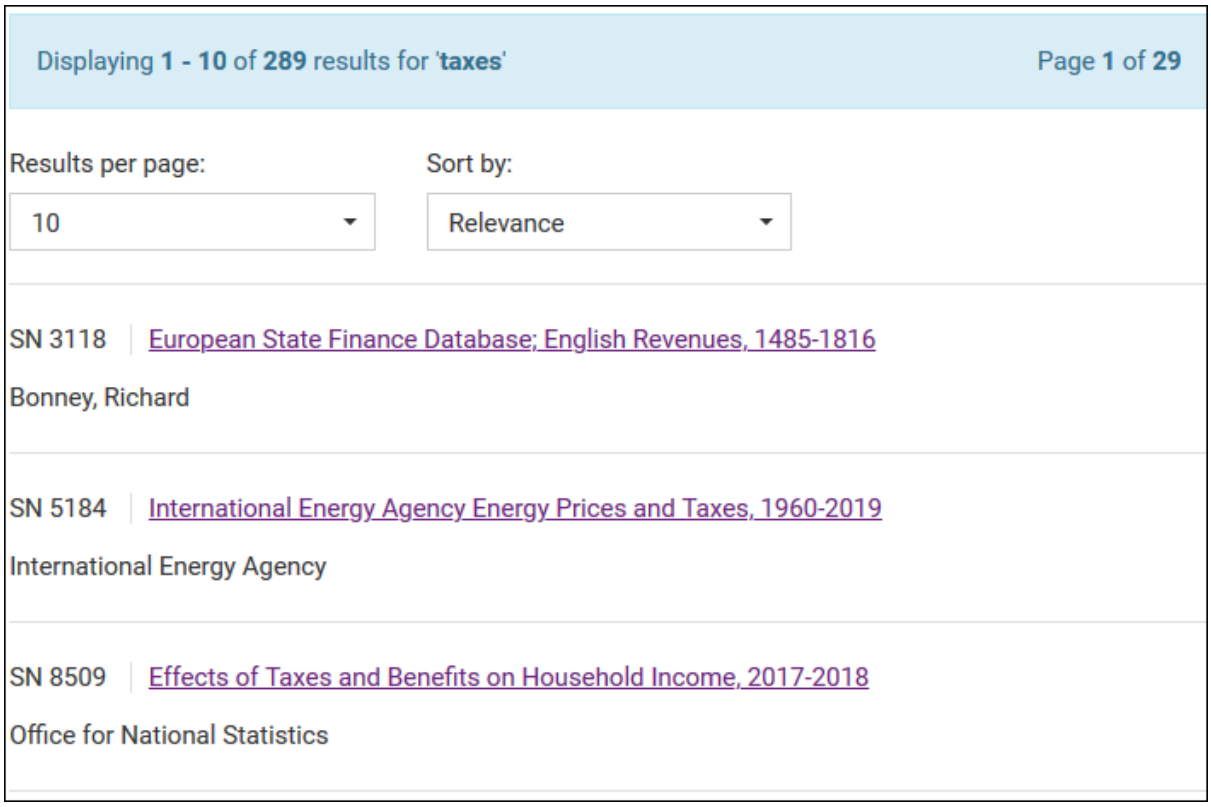

Figure 4.

Important tools such as electronic thesauri or ontologies and other types of controlled vocabularies and terminology provide an excellent extension and solution, for example, when a search for 'cat' returns studies about 'felines' (its broader concept), since the term 'cat' was not used for indexing but 'felines' was.

Improving findability and interoperability by implementing language-independent algorithms

Language-independent algorithms optimize the overall search to a great extent. An instruction such as 'search for ("x y") OR (x AND y)', applied to, for example, 'energy prices' searches automatically for *both* 'energy prices' (i.e. the words must be adjacent, but can vary in order) and 'energy' AND 'prices' (the words are not adjacent and can be in any order). The results list will display all studies for which the search terms are found at least once in the metadata record, more specifically, in the metadata fields in which the search is carried out.

Often, when results are found, they are displayed in a specific order according to, for example, 'Most recently released' or 'Relevance'. It is important to know which logic and algorithm sits behind the concept 'Relevance'. At UK Data Service 'relevance' is facilitated by using a score resulting from hits within a small, relevant set of available or populated metadata fields (see Table 1.), combined with a relative boosting weight (indicated in brackets) and a relative, proportional weight (a hit in a title of 3 words scores higher than a hit in a title of 10 words). In case of the same score, the results are ordered in descending order of version date. i.e. the most recent or newest first.

Title (50)	Country
Study number (15)	Geographical coverage
Abstract (10)	Spatial unit
Alternative title (10)	Town / village
Topic (10)	Other geography
Primary investigator (5)	Sampling procedure
Keyword (2)	Population
Data collector	Time period
Depositor	Time dimension
Sponsor	Kind of data
Grant number	Data type
Data producer	Language of study description
Series number	Language of study documentation
Subtitle	Data access tool
Type of key dataset	

Table 1. Metadata fields that are searched (with boost weight 50-1) for relevance ranking

Improving text mining by adopting tools based on computational linguistics

In this section, we focus on how linguistics, and more specifically computational linguistics, may contribute to and facilitate the mining and analysis of spoken and written data/material, i.e. qualitative data. For example, (pre-)processing interviews or oral histories for text-mining, i.e. preparing data for analysis and interpretation, may include the following steps and technology:

- Converting spoken data into written text data; parsing in order to group synonyms and multi-word expressions, sentence splitting; filtering by using stopword lists; discovering of patterns using frequency lists, concordances and correlations.
- Using automated speech recognition (ASR) tools, which are able to distinguish spoken natural language (sound) from surrounding noise. They are able to translate or convert spoken language into written language (spelling, text), which can be aligned with the spoken fragments. The transcriptions can also include, for example, repetitions, incomplete sentences or onomatopoeia (e.g. 'mmm', 'pfff').
- Adopting complementary or advanced technology, which is able to distinguish different speakers in interviews, i.e. speaker diarisation, or to tag positive/negative emotions, for example 'awful' (negative) vs. 'awfully nice' (positive). Multimodal technologies can also be used to investigate the importance of silences and role-taking in social interaction, tone and pitch, facial expressions and body language in audio-visual material.
- Enhanced speech-to-text transcription tools correctly assign capital letters - useful for the recognition of named entities and abbreviations -, sentence splitting and punctuation - useful for specifying questions or exclamations -, etc.

As we will demonstrate below, a wide range of Natural Language Processing (NLP) tools can indeed improve and offer help with the meaningful 'human' understanding or interpretation of texts.

Basic NLP can be adopted to enhance the quality of 'machine' generated frequency lists (which are currently not or only partially based on meaning/semantics), the creation of word clouds and term/keyword extraction.

Advanced NLP, such as syntactic parsers, help to disambiguate homophones like 'friar, fryer'; 'none, nun'; 'knead, kneed, need'; 'cense, cents, scents, sense'. They also detect co-referencing, identifying for example whether something refers to the same person or not, as in, for example, '*Warren* arrived early this evening. *The presidential candidate* was accompanied by *her* daughter'. This type of information is important when mining and analysing texts using frequency lists (based on both *meaning* and form) or investigating and focusing on one and the same person. It also contributes to the reduction of disclosure risk, for example, re de-identification (direct) and anonymization (indirect).

Another type of NLP tools concerns information extraction. Named Entity Recognizers (NER), for example, recognize and classify named entities, such as person names, organisations, locations or geospatial terminology, percentages, quantities, dates etc. The automatically generated and produced lists can be very useful for social science research, since they can assist with both simplifying anonymization and checking on or controlling disclosure and de-identification.

Using extraction tools for the detection of keywords and controlled terminology, in this case social science terminology and jargon, may also improve the quality of human text mining and analysis, including (semi-automatic) indexing. As mentioned before, as an example, the relevant meaning of 'was' as used in the UK Data Service data catalogue is 'Wealth and Assets Survey', and not the verb 'was' (which it would be for linguistic research concerning auxiliary verbs or passive constructions); 'CLOSER' stands for a range of longitudinal studies; its meaning in everyday language 'more close' is rather irrelevant in this respect.

As previously mentioned, keywords can be used to identify and describe the content of a study; they also improve information retrieval in terms of precision and recall, where precision is the result of dividing the number of true positives by the sum of all positives, and recall is the result of dividing the number of true positives by the sum of true positives and false negatives.
(NB the 'Keyword' metadata field has boosting factor 2, see above)

The following examples illustrate keyword extraction and auto-summarisation applied to a text example from the UK Data Archive website (Abstract copyright UK Data Service and data collection copyright owner):

> *The Commercial Victimisation Survey (CVS) provides a source of information on crime and crime-related issues as they affect businesses in England and Wales. It provides additional detail on the extent of crime to be used alongside the other main sources of information on crime. These are the Crime Survey for England and Wales (CSEW) (formerly the British Crime Survey), which covers crimes against private individuals and households, and the Police Recorded Crime statistics, which cover crimes reported to the police. In common with the CSEW, the CVS also includes crimes that are not reported to the police. The Police Recorded Crime data tables are available from the GOV.UK website.*
>
> *The CVS was conducted in 1994, 2002, 2012, 2013, 2014, 2015, 2016 and 2017 (at present, the Archive only holds data from 2002 onwards) and the survey has been commissioned to run in 2018. Further information on the CVS, with links to findings by year, can also be found on the GOV.UK Crimes against businesses webpage*.
>
> Keywords: *crime, information, England, cvs, csew, wales*
> (http://keywordextraction.net/keyword-extractor)
>
> Summary: *The Commercial Victimisation Survey (CVS) provides a source of information on crime and crime-related issues as they affect businesses in England and Wales. These are the Crime Survey for England and Wales (CSEW) (formerly the British Crime Survey), which covers crimes against private individuals and households, and the Police Recorded Crime statistics, which cover crimes reported to the police.* (https://summarygenerator.com/)

A lot of NLP tools also generate concordances (KWiC/Keywords in Context) and correlations in a text or group of texts, i.e. text corpus. This helps the human user to detect possible links and (unexpected) patterns, for example between 'school', 'learning', 'teachers' and 'knives'. Below is an example produced by SketchEngine's 'Concordance' functionality, when searching for the word 'travel' in an interview with a black immigrant to the UK. The human user can easily detect that interviewee 1 travelled before, but interviewee 2 travelled for the first time abroad by boat and that she/he disembarked in Southampton. An important fact here is that the answers from Respondent 1

and Respondent 2 have been identified and separated (see figure 5), a process similar to speaker diarisation in audio recordings.

Left context	KWIC	Right context
and then to - I don't remember if it is Euston or what it is. </s><s> Respondent 2: Actually I	travelled	by boat myself - I land at Southampton
Respondent 1: Well, I did give myself ten years. </s><s> This is the last time I'm going to	travel	, you know. </s><s> Interviewer: You ha
s. </s><s> This is the last time I'm going to travel, you know. </s><s> Interviewer: You had	travelled	before then? </s><s> Respondent 1: Ye
d travelled before then? </s><s> Respondent 1: Yes. </s><s> Interviewer: Where had you	travelled	before to? </s><s> Respondent 1: In Ar
d. </s><s> Interviewer: Did you remember anything the government might have said about	travelling	at this time? </s><s> They want people
espondent 1: My.... </s><s> Respondent 2: Probably because it was my first - well, my frist	travelling	abroad. </s><s> I don't think it is very ea

Figure 5.

The Voyant OS online tool, for example, offers a 'Correlations' functionality. Correlations are words or terminology that often appear in each other's neighbourhood. Searching for 'legal' in a text about medicinal cannabis in California informs the human user that both the laws for the use of cannabis and the attitudes towards it have changed, since it became legal in the 1990s (see figure 6.).

Contexts | Bubblelines | Correlations

Term 1	←	→	Term 2	Correlation (r)	Significanc...
cannabis			legal*	0.70515615	0.022741713
became			legal*	0.41442487	0.233756
early			legal*	0.41442487	0.23375599
laws			legal*	0.3563483	0.31215996
1990s			legal*	0.13363063	0.71284723
attitudes			legal*	0.12982272	0.7207509
1980s			legal*	0.102062084	0.77905035
changed			legal*	0.102062084	0.77905035
largely			legal*	0.102062084	0.77905035

legal* × 20 minimum coverage (%100)

Figure 6.

Concluding remark

All of the tools and services described above are available for different languages. It is certainly worth considering to add these linguistics-based services and tools to the standard package of

 Beeken, Jeannine (2020) A recommendation to the SSH community: Take a linguist on board, IASSIST Quarterly 45(1), pp. 1-10. DOI: https://doi.org/10.29173/iq992

services and infrastructure offered by archives to social science and humanities researchers, because they

- Promote and support multidisciplinary research and cooperation.
- Facilitate interoperability between research approaches and methods, technology and tools.
- Increase awareness of a wide variety of language technology tools which may assist or improve SSH research.
- Illustrate and demonstrate the potential and benefit of computational social science.
- Result in a better user experience with search, retrieval, extraction and analysis tools and create a better understanding, and therefore openness to unknown or lesser-known technology.

References

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: ten years on. *Lexicography*, 1: 7-36, 2014. [BibTeX] [Download PDF]. http://www.sketchengine.eu

Home Office, Crime and Policing Analysis Unit. (2020). *Commercial Victimisation Survey, 2017*. [data collection]. UK Data Service. SN: 8352, http://doi.org/10.5255/UKDA-SN-8352-1

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web. http://voyant-tools.org/.

End-notes

[1] Jeannine Beeken is Senior Metadata and Ontologies Officer at the UK Data Service, University of Essex, UK. She can be reached by email: Jeannine.beeken@essex.ac.uk