JSI Newsfeed Corpus

Jan Bušta (Lexical Computing Ltd, Czech Republic), Ondřej Herman (Masaryk University, Czech Republic), Miloš Jakubíček (Lexical Computing Ltd, Czech Republic), Simon Krek (Institut Jožef Stefan, Slovenia) and Blaž Novak (Institut Jožef Stefan, Slovenia)

The JSI Newsfeed corpus is a new family of web corpora created from the JSI newsfeed created by Jozef Stefan Institute, Slovenia (Trampus et al 2004). JSI newsfeed is a clean, continuous, real-time aggregated stream of semantically enriched news articles from RSS-enabled sites across the world. The newsfeed is available in many languages including Arabic (250 million words), Catalan (40 million), Czech (146 million), German (936 million), English (8 billion), Finnish (75 million), French (782 million), Croatian (150 million), Hungarian (77 million), Italian (335 million), Korean (130 million), Dutch (176 million), Polish (150 million), Russian (500 million), Spanish (1.5 billion), Serbian (38 million), Swedish (147 million). The project continuously processes 80,000 RSS feeds which bring between 350,000 and 600,000 articles every day. All articles are cleaned so that the main body of text is included, duplicated articles are removed and, most importantly, all data is time stamped.

The JSI newsfeed has been processed into the JSI Newsfeed corpus by Sketch Engine (Kilgarriff 2014). Sketch Engine is a corpus query software with automated corpus building tools. Sketch Engine currently holds 150 TB of data including 400 corpora in 85+ languages. Sketch Engine specializes in processing extremely large corpora with a size of up to dozens of billions of words.

The JSI Newsfeed corpus was tagged for parts of speech and the time stamps were used to augment the corpus with diachronic annotation. Currently, the corpus covers the time period from the year 2014 onwards. The feeds are crawled continuously by Jozef Stefan Institute and the corpus is amended daily with the latest articles. By combining this data with other English web feed corpora, a total period from 2009 up to the current day is covered.

The diachronic annotation is extremely valuable in connection with Sketch Engine and its trends feature. The trends feature analyses the frequency of the use of a word in time by comparing the frequency of use across a series of comparable time periods.

The algorithms will then present the user with a list of words which have changed its frequency of use over a longer time period, i.e. neologisms candidates and words going out of use.

word	Trend		p-value	Freq	Graph
recordkeeper	1.3763	+	0.005107	194	
mountainbike	-1.3763	−	0.007049	158	
kanyewest	1.3763	+	0.034572	245	
pero	-1.3763	−	0.001524	1,022	
hindi	-1.3763	−	0.000176	2,433	
ois	-1.3763	−	0.000873	384	
partypoker	-1.3763	−	0.020090	163	
playdowns	-1.3763	−	0.001650	254	
rootworms	-1.3763	−	0.002528	144	
biosimilar	1.3763	+	0.000023	4,562	
digitalisation	1.3763	+	0.000005	1,013	
barangay	-1.3763	−	0.002181	4,268	
remasters	1.3763	+	0.000037	722	
nonnegative	1.3763	+	0.000400	512	
deconvolution	1.3763	+	0.004330	222	
wd	-1.3763	−	0.007049	307	
numismatics	-1.3763	−	0.000488	319	
placegetters	1.3763	+	0.008250	238	
megaloads	-1.3763	−	0.000204	160	
northerns	-1.3763	−	0.001974	182	
integrable	1.3763	+	0.000873	969	
strawweight	1.3763	+	0.000176	1,416	

The words listed on the results screen may contain certain items which do not qualify as words whose use changed over time, typically, misspelt words might fall into this category. This, however, does not reduce the usefulness of such a tool. Even with a certain level of pollution, it will take a lexicographer only fraction of time to discover neologisms compared to the traditional approach of reading media trying to discover a

new word. The results are sorted by the trend value, i.e. the words with the biggest change are at the top irrespective of whether the change was positive (=growing usage) or negative (=decreasing usage).

In Sketch Engine, the trend value can be calculated using two methods: either a simple linear regression is used, or the Mann-Kendall test. The latter is used because it showed to be more robust with noisy data, such as very imbalanced time slices, or random local extremes, both of which occur very often especially with web corpora. Methods of calculating trends are discussed in great detail in (Kilgarriff, 2015) and (Herman, 2013).

Apart from the trend value, the p-value is calculated and shown, and for each item in the list the interface allows the user to show concordance lines and frequency distribution according to the time periods.

The trends feature on its own cannot be used to identify changes in word senses. The current research focuses on exploiting the diachronic annotation even further in the direction of an automatic identification of the semantic shift. The approach combines word sketches, one-page summaries of a word's collocational behaviour, with the trends feature. Preliminary research (Baisa, 2015) shows that attention should be paid to the change of the collocational behaviour of the word in time because changes in collocational behaviour correlate with word sense change. This could lead to a new functionality in Sketch Engine aimed at automatic identification of the semantic shift with the help of diachronically annotated corpora.

References

Baisa, Herman, & Jakubíček (2015). Towards Automatic Finding of Word Sense Changes in Time. *RASLAN 2015 Recent Advances in Slavonic Natural Language Processing*, 33.

Herman & Kovář (2013). Methods for detection of word usage over time. *RASLAN 2013 Recent Advances in Slavonic Natural Language Processing*, 79.

Kilgarriff, Herman, Bušta, Rychlý, & Jakubíček (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics* (pp. 65-70).

Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, & Suchomel (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Trampuš & Novak (2012). Internals of an aggregated web news feed. In *Proceedings of the Fifteenth International Information Science Conference IS SiKDD 2012* (pp. 431-434).