

Automating Stable Lexical Marker Analysis with webtool SnelSLiM

Currently, the most convenient way to perform Stable Lexical Marker Analysis is using the mclm R package developed by Prof. Dr. Dirk Speelman. But even this package has only recently become available on <https://wwwling.arts.kuleuven.be/mclm/> and is not very well known yet. This leaves linguists worldwide with an interest in keyword analysis but no tool to perform it with. SnelSLiM aims to solve this issue by making SLMA simple and quick.

What is SLMA?

SLMA, or Stable Lexical Marker Analysis, is a form of keyword analysis developed initially by Speelman et al. 2008 and further refined in De Hertog et al. 2012. It uses the standard occurrence table also used in collocational and collostructional analysis.

	keyword	other words
Selected Fragment	a	b
Contrasted Fragment	c	d

Then a G test (also known as G squared or log-likelihood) is used to check for significant keyword-ness, while log odds ratio is used to measure effect size. This is done for every combination of fragments of both corpora.

Features of snelSLiM

- ▶ Wide range of supported corpus formats including FoLiA, DCOI and TEI/BNC
- ▶ User-defined XPath queries for any XML format
- ▶ Extremely fast and efficient binary component
- ▶ Can be deployed on generic, cheap web hosting
- ▶ Support for accounts and saved corpora
- ▶ Report results are generated on the server and always online available
- ▶ Under development: n-gram support, collocational analysis based on SLMA results, university login integration. Suggestions welcome!

What is snelSLiM?

SnelSLiM is a webtool to quickly and easily perform SLMA without requiring any knowledge of R or programming. I started developing the application as my bachelor paper and have continued development for my master thesis under the supervision of Prof. Dr. Dirk Speelman.

Main interface of snelSLiM

#	Marker	Absolute score	Normalized score	Attraction	Repulsion	Lowest Log Odds Ratio	Highest Log Odds Ratio	StdDev	Log Odds Ratio Score
1	x	67125	0.255648709778854	67704	620	-16.83283103051837	18.56137772512026	4.6373882388873	3.0767858238339
2	xx	36425	0.13858942481122	37282	837	-18.05411875616696	20.40482723209028	4.8787928222499	1.887072889315424
3	xxx	30552	0.1188589458553823	31081	429	-18.28901070819124	20.48103734946507	3.883133772979555	1.613814918552458
4	greet	32519	0.1236476558491806	32799	280	-16.148801191778855	19.79458880431459	4.22136138866455	1.514773094484797
5	has	25821	0.0984174537471276	26752	931	-18.05411875616696	19.79458880431459	5.72732852464673	1.263074623877059
6	is	27211	0.103715470880769	29126	1515	-16.920823132738	17.460746050521297	6.80880828545382	1.37445581352347
7	mean	28316	0.1079272108805546	30116	1800	-15.254880578733078	16.67077011028386	6.44820276124946	0.945735927828182
8	not	28463	0.1054861154891217	30412	1749	-16.86241118803066	16.808787018100148	6.511148423880942	0.945735927828182
9	even	21580	0.0825257626328231	23451	1871	-20.761633094818784	17.982175781786	7.29401042074823	0.882555515543149
10	do	18531	0.064523273231294	22091	5140	-18.38411385240635	19.79536795022688	10.42851345421035	0.827973781978883
11	think	28534	0.0748088588857948	21235	701	-14.537480435762784	16.4797148773021234	5.887878814231481	0.7887744212285407
12	but	17743	0.061927933923781	18861	1958	-16.975446704337002	18.405680880830993	6.12288871139719	0.779548843183382
13	vanavond	17882	0.06515778257373	19079	1697	-16.20945482571422	17.02776178789531	7.13668123742955	0.750208922889414
14	ifagen	18009	0.06101874500258234	17236	1227	-16.079879787156444	17.02217713286485	6.428868233308105	0.76078795957807
15	but	17124	0.061927933923781	18861	1958	-16.975446704337002	18.405680880830993	6.12288871139719	0.779548843183382

A snelSLiM report comparing the SoNaR New Media SMS subcorpus with a subset of the tweet and chat subcorpora of the same corpus.

Please feel free to give me feedback or request a demo.

Visit <http://tiny.cc/snelSlim> for more details.