Scientific Report of Short Term Scientific Mission

COST STSM Reference Number: COST-STSM-IS1305- 24203 Period: 25-01-2015 to 07-03-2015 Duration: 30 working days COST Action: IS1305

STSM type: Regular (from Slovenia to Poland)
STSM Title: Automatic Extraction of Polish Multi-Word Expressions
Guest/STSM applicant: Cyprian Laskowski, University of Ljubljana
Host: Prof. Piotr Żmigrodzki, Instytut Języka Polskiego, <u>piotr@ijp-pan.krakow.pl</u>

1. Purpose of the STSM

The main goal of my STSM was to take methods used in Slovenia for automatically extracting multiword expressions (MWEs), apply them to Polish and demonstrate the usefulness of this approach at the Institute of the Polish Language (IPL) in Kraków, where the Great Dictionary of the Polish Language (GDPL) is compiled and edited. A secondary and more ambitious goal was to visit the Linguistic Engineering Group (LEG) in Warsaw and the plWordNet group (PWNG) in Wrocław and apply automatic methods for assessing the degree of semantic compositionality to the multi-word expressions I had extracted.

2. Description of the work carried out during the STSM

I started and ended my trip at the IPL. On the first day, I was warmly welcomed by Prof. Żmigrodzki, introduced my project goals and Sketch Engine to the GDPL staff, and had several discussions with several staff members to understand their current tools, methods and perspectives for handling MWEs. Then I set up my Sketch Engine account for the work, selecting a testing corpus already available on Sketch Engine which was of an appropriate size and which used the same tagset as the National Corpus of Polish (NCP). I then engaged the bulk of the work, preparing the Polish sketch grammar.

I started with the well-established Slovene sketch grammar, which was developed by Dr. Krek and colleagues in Slovenia a few years ago, and transformed the Slovene sketch grammar for Polish. This involved mainly 3 parts. First, I changed the regular expressions for matching elements in the tagged corpus to fit the tagset used my NCP (e.g., [tag="(subst|depr|ger):.*"] for collocate nouns). Second, I adjusted the gramrels to reflect the differences between Polish and Slovene grammar (e.g., adjectives come before nouns in Slovene, but can come before or after nouns in Polish). Third, I systematically renamed the macros and gramrels (grammatical relations) for Polish (e.g., rzecz_d_1" for head nouns in genitive (dopełniacz) case). I iteratively made changes, uploaded and recompiled

the changed sketch grammar, and tested the results on a few words in the Sketch Engine interface. I also adjusted the Python script provided by Sketch Engine for batch extraction and the related configuration files for Polish and used the script to extract the collocations for a few words in XML format. I had a few conceptual and technical issues along the way, but I was in touch with my Slovene colleagues about pecularities of the Slovene sketch grammar and the Sketch Engine support staff about technical difficulties.

Once I had more or less prepared the Polish sketch grammar, I presented the interim results to the GDPL staff, showing them a few gramrels along with the tabulated results in the Sketch Engine interface for a few words. I got some useful feedback concerning a few gramrels, and was asked if I could also adjust the organisation of the gramrels to that of their editorial system. This effectively meant restructuring and renaming my gramrels to be in a 1-to-1 relationship with their collocation categories. As they had given me access to their dictionary editing pages, I then studied these categories more closely and reorganised my gramrels accordingly. To that end, it was very useful that the desk and space they provided me with was in the same room as two of their dictionary editors, who clarified various details and problematic cases for me during my stay. Some of these turned out to be rather subtle. For instance, gerunds are generally treated as verbs, so that the collocates of gerunds were listed under the corresponding verb entry, but gerund forms were included among the noun collocates of adjectives. This meant that the final gramrels were not as simple and symmetric as the initial versions.

At this point, I went to Warsaw and Wrocław to address the other part of my STSM, which was to attempt to automatically detect the semantic compositionality of collocations. When I came to Warsaw, I met with Prof. Przepiórkowski, with whom I had arranged to come and work on this problem during this period. Unfortunately, as he had warned me, this was a very busy period for him, and in fact even busier than he had expected. Also, it turned out when we talked that although his group was interested in this problem, they had not yet developed tools or methods to address it. Therefore, although I presented my STSM goals to his group and also attended an interesting presentation on the valience dictionary they have been developing, I quickly switched my focus to Wrocław. I contacted Prof. Piasecki's team and asked them about their approach to the problem. They were very helpful and communicative, so I spent much of my week in Warsaw familiarising myself with their methods remotely.

Then I went to Wrocław and communicated more closely with Prof. Piasecki's enthusiastic PWNG team. The basis of their approach is to statistically compare the distribution of words with the distributions of MWEs including those words. They look not only at text windows, but at grammatical relations that the words or MWEs enter into with specific other words. The basic idea is that if a MWE is semantically compositional, then the distribution of the head word is generally more likely to be similar to the distribution of the entire MWE. They also used additional heuristics in combination with this method, such as whether the word order of a MWE is fixed and whether it can have other words inserted in between. Their studies had shown that the method they were developing for nounadjective collocations was already achieving good results. I therefore wanted to try to apply their tools to my data to get semantic compositionality values for a small sample of my noun-adjective MWEs.

Unfortunately, this turned out to be too ambitious a task to do in such a short period of time. First of all, their method relies on having data for several grammatical relations for both the headwords and the collocations on a large corpus. As I had only been using a small corpus, I had to make a simplified version of my sketch grammar and fetch the data on a large corpus. I got permission from the Sketch Engine staff to use their largest corpus, but extracting these results took several days. The PWNG did offer me to use their own corpus data, but I felt I should use the data I had gathered; in hindsight, this was probably a mistake. Second, I needed to convert the data from Sketch Engine to a suitable

format for the PWNG's SuperMatrix tool, but as I had not done this before, I made a subtle but significant mistake in the conversion, which greatly reduced the validity of my data. Unfortunately, by the time Prof. Piasecki noticed and informed me of my mistake, the data had already been mostly processed by SuperMatrix and there wasn't time to do it again. Third, in my attempt to select interesting nouns and collocations to test, I had selected mostly head nouns which were polysemic, which reduces the reliability of their method. Fourth, again due to time limitations, I wasn't able to incorporate the other heuristics mentioned above(word order, separability) into the method. Fifth, as I had only wanted to make a small demonstration and was short on time, I had only selected a small number of collocations to test, but this did not allow for a proper statistical analysis of the results, which was particularly problematic given the other compromises and limitations listed above. As a result, my results were not adequate and I have to defer this ambitious goal to the future. My experience in Wrocław taught me that this problem can be effectively tackled in clever ways, and that the PWNG team is wonderful to work with, but that this is indeed a difficult problem and cannot be solved without the proper considerations and time.

After my stay in Wrocław, I returned to Kraków. I did a few more tests and improvements of my Polish sketch grammar. Then I thought about how I could most effectively show them its potential usefulness, and decided to do so directly in the context of their internal website. I mapped the XML collocations downloaded for a sample word to JavaScript format, and added them to a saved copy of their editor page for a particular word. I also added an HTML button on this page, which, when clicked, showed the list of downloaded MWEs at the top of the window, along with checkboxes and buttons, which would allow them to select which meanings of the word the MWE should be assigned to and buttons for pasting the word directly in the appropriate MWE textarea (based on the 1-to-1 mapping between gramrel and collocation category). I implemented this partially for one example, enough to show how automatically fetched collocations could be usefully integrated into their existing internal editor webpages.

Finally, I gave a concluding presentation and demo. I showed the final sketch grammar, along with a few examples, and one extended example showing how my sketch grammar found most of the collocations they had listed in their dictionary for a common word, even with my small testing corpus. Then I showed them my edited version of their editor page. The response was enthusiastic, especially once they saw their edited webpage and saw concretely how such methods could help them directly with their daily work. Prof. Żmigrodzki said that I have given them a lot to think about.

In all, I think the main goal of my STSM was very successful: I developed a Polish sketch grammar adapted for the GDPL group, showed them examples of extracted MWE results, and demonstrated how this approach could be directly integrated into their daily work. And although the more ambitious goal of automatic computation of the semantic compositionality of collocations was less successful, my appetite has been whetted and I hope to pursue this problem further soon, hopefully in further collaboration with the PWNG. Also, I hope that by visiting and discussing with three different leading lexicography-related research groups in Poland, I have also helped them develop potential further colloboration with each other. In any case, I have certainly learned a lot.

I include below screenshots of my Sketch Engine results for an example word ("samochód" = "car"). I attach also my sketch grammar and an example (also for "samochód") of the collocations extracted in XML with the Sketch Engine Script. I do not include the adjusted HTML editor page, as this is internal to GDPL.

														•	Q Send feedb	ack o	orpus: <u>plTenTen12</u>	[sample] Dr. C
am	noch	ód	(noun piTen) Ten12 [sample] freq	= <u>13,183</u>	(274.1	8 per	million)										
rzecz n	ad przym	8,412	1.80	rzecz pod rzecz o	4.811	2.40	[Ze	rz pod czas podm	1,532	1.60	rzecz pod czas t	1,233	2.40	IZes	cz nad rzecz d	912	0.50	
🗌 cięż	tarowy	413	10.49	wypożyczalnia	178	9.85		zderzyć	20	8.88	parkować	21	8.54	0	marka	249	9.08	
osol	bowy	790	10.34	wynajem	122	9.26	0	wjechać	12	7.53	zaparkować	19	8.52	8	volvo	2	7.72	
dost	tawczy	228	9.69	🗐 naprawa	66	7.93	0	uderzyć	8	6.41	🗐 naprawtać	13	7.93		BMW	10	7.60	
tere	nowy	131	8.64	🗐 zakup	182	7.75	曰	jeżdzić	11	6.21	🗐 wynająć	26	7.73	0	Suzuki	6	7.31	
služ	tbowy	101	8.35	kupno kupno	48	7.66	0	jechać	13	5.96	🗐 kupić	60	7.72	0	ford	6	7.13	
uży	wać	194	7.80	zderzenie	32	7.66		kosztować	Z	5.83	kupować	22	7.65	63	segment	13	6.83	
E luks	usowy	68	7.77	kierowca	83	7.64		zatrzymać	11	5.78	🖂 zostawić	19	7.62		ratownictwo	8	6.63	
Zast	epczy	56	7.49	parkować	32	7.59	0	stać	36	5.37	ukraść	8	7.37	10	Leszek	6	5.88	
zaby	ytkowy	56	7.36	wypożyczyć	35	7.50	8	clesać	6	5.23	wypożyczyć	12	7.18	0	klasa	36	5.41	
🗆 zapi	arkować	34	6.97	🔲 oklejać	27	7.45	63	sprawlać	2	5.03	zatrzymywać	11	7.04	0	straž	10	5.40	
elek	tryczny	58	6.88	🖯 kradzież	23	7.36		pojawić	13	4.63	🗐 wjechać	Z	6.97	E.	typ	33	4.52	
uszk	kodzony	33	6.88	model	134	7.32	6	sprawdzać	6	4.58	zostawiać	9	6.95	12	świat	9	3.56	
spor	rtowy	Z2	6.71	producent	109	7.29		kupować	6	4.49	wyprzedzać	6	6.69	8	uczestnik	6	Z.41	
rajd	lowy	29	6.71	🗐 skup	26	7.26	8	poruszać	6	4.45	🗐 sprzedać	13	6.46	8	rok	30	1.70	
🗌 wyś	clgowy	26	6.57	🔲 sprzedaż	134	7.20		musieć	29	4.34	🗐 zakuptć	18	6.38	8	firma	Z	0.75	
now	vy	323	6.56	postadacz	35	7.15	<u>E</u>	stawać	8	4.29	abywać	<u>B</u>	6.30	0	osoba	Z	0.55	
wta:	sny	132	6.54	dealer	24	7.14		potrafić	Z	4.14	🗉 myć	0	6.26				>>	
drog	gi	38	6.41	🗐 flota	27	7.06	0	posladać	24	4.11	zatrzymać	12	5.97	-		-		
użyt	tkowy	29	6.35	🗐 fabryka	32	7.03	E	uczestniczyć	Z	4.02	produkować	13	5.80					
twó	1	94	6.24	zaparkować	19	6.88		czekać	6	3.97	🖾 pozostawić	Z	5.62					
pryv	watny	43	6.20	karoserta	18	6.87	5	być	425	3.93	jeżdzić	6	5.42					
🗐 firm	lowy	27	6.19	właściciel	20	6.82	8	zostać	22	3.72	sprzedawać	8	5.29					
nape	ędzać	20	6.13	ubezpieczenie	64	6.79	123	wyglądać	6	3.50	postadać	44	5.00					
poże	amiczy	19	6.13	leasing	21	6.79	8	znajdować	13	3.48	🗉 nabyć	5	4.98					
kem	pingowy	18	6.09	🔲 myć	19	6.66	0	matć	14	3.38	prowadzić	63	4.80					
			>>			>>				>>			>>					
_			hereinend	u		Conversed.	-		_				-					

opel 20 8.41 jack 21 0.2	opel 20 8.41 jazda 22 6.42 9.22 1000 7.44 jechzik 24 9.42 1 royota 1 8.53 jechzik 2 9.64 jechzik 2 9.64 1 royota 1 8.53 jechzik 2 9.64 jechzik 2 9.64 1 royota 1 8.57 jopotróż 4.64 9.73 jopotróż 9.74 jopotróż 9.74 jopotróż 9.74 jopotróż 9.74 jopotróż 10.64 jopotróż 10.64 jopotróż 10.64 jopotróż 10.64 jopotróż 10.64 jopotróż 10.77 jopotróż 9.77 jopotróż jopotróż 10.77 jopotróż 10.77 jopotróż jopotróż 10.77 jopotróż jopotróż 10.77 jopotróż <	FZPCZ pad FZP	z m 81	7 1.1	0 77	ecz pod rzecz r	693	10.30	1200	z nad Dez	628	2.20	170	ez pod czas n	377	4.90	57	ered	393	0.60		
Image: Solution of the second of the seco	ford 15 8.55 dojacd 10 7.79 ford 15 8.57 przejadźka 9 7.67 300 2 6.41 dojacd 8 8.71 rycyta 14 8.77 połcić 45 6.77 500 6 6.13 przejechać 9 7.77 atath 8 8.11 połcić 45 6.77 500 6 5.78 6 9 7.77 velkawgen 17 7.91 bandel 19 5.58 600 5.58 6 900 5.54 900 6.43 połczić 5 7.07 welkawgen 17 7.91 bandel 19 5.54 900 900 10 7.07 atath 13 7.54 transport 19 5.57 900 900 10 7.07 filt 13 7.77 metzkawe 13 5.51 900 10 17 7.07 star 9 7.78 metzkawe 10 10 11 <t< th=""><th>opel</th><th></th><th>0 8.</th><th>1</th><th>jazda</th><th>238</th><th>9.25</th><th></th><th>1000</th><th>7</th><th>6.84</th><th>1</th><th>jeżdzić</th><th>96</th><th>9.65</th><th>10</th><th>motocykl</th><th>39</th><th>9.22</th><th></th><th></th></t<>	opel		0 8.	1	jazda	238	9.25		1000	7	6.84	1	jeżdzić	96	9.65	10	motocykl	39	9.22		
1 Toyota 1 8.27 prepatzižka 9 7.47 300 2 6.41 dogechać 19 8.22 samolac 9 5.70 volvo 12 8.24 volzavad 8 6.77 500 6.13 projechać 9 7.71 projechać 9 7.72 projechać 9 7.72 projechać 9 11 1.74 1 7.72 projechać 9	1 optola 1 8.27 0 0.00 7 6.41 0 0 2 8.12 1 volvo 12 8.24 0 wjazd 8 6.77 10 0 5 8.81 0 0 5 7.77 1 0 0 5 5.86 0 0 5 8.81 0 0 5 9.84 0 0 5 8.41 0 0 5 9.84 0 0 0 10 7.77 0 0 5 5.84 0	ford		5 8.	5	dojazd	30	7.94		dwa	135	6.71		Jechać	Z1	8.63	0	tramwaj	12	7.61		
• vylov 12 8.24 • vydav 1 6.47 • try 11 6.13 • podráčovač 22 8.04 • autobu § 5.33 • volkswagen 11 7.41 6.13 9 podráčovač 6 7.37 • podráčovač 7.37 • podráčovač 6 7.47 • podráčovač 13 7.07 • podráčovač 6 4.47 • aud 10 7.79 • mintuta 8 4.54 • crtery 13 5.46 • dyponovač 1 6 4.43 • frad 2 7.73 • mieczkani 6 4.43 • dyponovač 1 4.44 2 5.46 6 4.44 • • 0 0	• vboo 12 82.42 • vjad 8 6.97 • try 41 6.13 • podráce 22 8.04 • strobus 6 5.33 • volkovagen 11 7.97 500 6 6.13 9 7.37 • volkovagen 11 7.78 • handel 10 5.46 9 7.37 • maxima 8 4.50 • audi 10 7.77 • maxima 8 4.54 • creery 11 5.46 9/poproval 21 6.41 • audi 10 7.77 • mylazd 5 1.77 porticals 5.10 maxima 8 4.50 • fist 12 7.77 • mylazd 1.12 13 2.53 9 dotata 6 6.14 • ratowicco 7.77 • mylazd 1.14 13 2.53 9 dotata 6 4.11 • ratowicco 7.77 • mylazd 2 5.17 5 12 5.37 9 dotata 6 4.14 >>> >>>	Toyota		4 8.	7	przejażdźka	9	7.67		300	Z	6.41		dojechać	18	8.32	0	samolot	9	5.70		
abath 8 8.13 porki 9 6 6.77 500 6 6.13 porkietat 9 7.37 prewer 2 5.13 scion 6 7.87 1 6.64 10 5.64 0 0 6 5.74 0 porkietat 9 7.37 1 prewer 2 5.13 add 10 7.77 10 10 3.64 0 2.07 3.27 10 maxsyna 2 4.50 filat 10 7.77 markyna 2 3.45 13 5.45 0 0 0 0 maxsyna 2 4.50 filat 12 7.77 markyna 3 3.37 13 12 5.51 12 5.51 0 0 0 0 10 <td>a whith 8 8.13 przejazd 14 6.36 80 6 6.13 przejechać 9 7.37 szton 7.65 handel 12 5.46 80 6 5.76 pociąg 6 5.10 marek 18 7.44 Utranport 10 4.56 5.96 porzyschać 12 7.07 marek 18 7.77 uminuta 4.45 vielswagen 17.77 kerować 5 4.41 fat 10 7.77 uminuta 4.54 vielswagen 6 5.31 dogadzać 6 5.34 frat 12 7.77 uminuta 4.51 vielswagen 6 4.63 ratowniczo 2 7.70 umatoz 1 1.44 12 2.31 popichác 5 0.33 star 7.73 umatoz 1 1.44 12 2.31 popichác 5 0.33 gada 1 1.44 2 2.12 3.23 dotrzeć 4.43 4.43 1.44</td> <td>🔲 votvo</td> <td></td> <td>2 8.</td> <td>4</td> <td>wjazd</td> <td>8</td> <td>6.97</td> <td>8</td> <td>trzy</td> <td>41</td> <td>6.15</td> <td></td> <td>podróżować</td> <td>27</td> <td>8.04</td> <td>0</td> <td>autobus</td> <td>6</td> <td>5.33</td> <td></td> <td></td>	a whith 8 8.13 przejazd 14 6.36 80 6 6.13 przejechać 9 7.37 szton 7.65 handel 12 5.46 80 6 5.76 pociąg 6 5.10 marek 18 7.44 Utranport 10 4.56 5.96 porzyschać 12 7.07 marek 18 7.77 uminuta 4.45 vielswagen 17.77 kerować 5 4.41 fat 10 7.77 uminuta 4.54 vielswagen 6 5.31 dogadzać 6 5.34 frat 12 7.77 uminuta 4.51 vielswagen 6 4.63 ratowniczo 2 7.70 umatoz 1 1.44 12 2.31 popichác 5 0.33 star 7.73 umatoz 1 1.44 12 2.31 popichác 5 0.33 gada 1 1.44 2 2.12 3.23 dotrzeć 4.43 4.43 1.44	🔲 votvo		2 8.	4	wjazd	8	6.97	8	trzy	41	6.15		podróżować	27	8.04	0	autobus	6	5.33		
• volkswagen 11 7.91 0 10 7.64 0	1 10 6 5.86 0 dojestate 5 7.33 0 pociag 6 5.00 1 10 7.75 1 handel 10 5.64 200 6 5.86 0 pociag 6 5.00 1 10 7.75 1 handel 10 5.64 5.77 0 pociag 6 3.00 1 10 7.79 10 4.43 0 pociag 6 3.13 1 11 7.77 1 11 5.17 0 poletad 6 6.14 1 1 7.77 1 11 2.71 1 5 12 5.07 0 poletad 6 6.14 1 1 10 7.70 11 2.71 1 4 15 2.11 0 dotract 6 4.41 1 1 10 1.0 1 2.01 0 dotract 6 4.41 1 1 1 1 2.17 1 4 15 2.11 0 0 dotract 6 4.41 0	abarth		8 8.	3	podróż	48	6.77		500	6	6.13	101	przejechać	9	7.37	13	rower	Z	5.13		
scon 9 7.83 1 hadel 10 2.64 200 2 3.64 10 7.77 1 margka 8 4.50 audi 10 7.79 1 transport 1 5.77 0 opparation 2 7.70 1 imargka 8 4.50 filat 12 7.77 1 minuta 8 4.54 1 wr(c) 22 5.60 1 6.94 1 minuta 8 4.57 1 wr(c) 22 5.60 1 6.94 1 1 10 10 2.24 5.60 1 10 10 2.24 1 10 </td <td>scien 9 7.83 handel 19 5.64 000 \$ 5.84 prorpiechal 13 7.07 hassing a 8 4.50 add 10 7.73 transport 19 4.54 transport 13 5.64 dyporound 13 7.07 histowca 6 4.40 filt 13 7.73 minuta 8 4.54 cracery 13 5.64 histowca 14 6.54 filt 13 7.73 droga 11 2.71 5 5.73 dodtal 6 4.63 recault 9 7.72 droga 11 2.71 5 5.73 dodtal 6 4.63 star 2 7.58 osoba 11 1.49 16 5 5.07</td> <td>volkswage</td> <td>61 24</td> <td>1 7.</td> <td>1</td> <td>przejazd</td> <td>14</td> <td>6.36</td> <td>8</td> <td>80</td> <td>6</td> <td>5.98</td> <td>0</td> <td>dojeżdżać</td> <td>6</td> <td>7.33</td> <td>10</td> <td>pociag</td> <td>6</td> <td>5.10</td> <td></td> <td></td>	scien 9 7.83 handel 19 5.64 000 \$ 5.84 prorpiechal 13 7.07 hassing a 8 4.50 add 10 7.73 transport 19 4.54 transport 13 5.64 dyporound 13 7.07 histowca 6 4.40 filt 13 7.73 minuta 8 4.54 cracery 13 5.64 histowca 14 6.54 filt 13 7.73 droga 11 2.71 5 5.73 dodtal 6 4.63 recault 9 7.72 droga 11 2.71 5 5.73 dodtal 6 4.63 star 2 7.58 osoba 11 1.49 16 5 5.07	volkswage	61 24	1 7.	1	przejazd	14	6.36	8	80	6	5.98	0	dojeżdżać	6	7.33	10	pociag	6	5.10		
marek 19 7.24 10 10 7.63 10 7.02 10 901232 2.2 7.01 10 minuta 6 4.54 10 0prované 21 6.94 10 minuta 6 4.54 10 0prované 21 6.94 10 minuta 6 3.13 13 2 5.54 0prované 21 6.94 10 0mr 10 2.2.74 10 0mr 10	mark 19 7.74 transport 10 4.58	scton		g 7.	13	handel	10	3.64	6	200	\$	5.84	0	przyjechać	13	7.07	0	maszyna	ā	4.50		
audi 10 7.79 imituta 1 4.54 i creary 1.1 5.46 i dyponovací 2 1 6.46 initat 1 7.77 i mglczyma 1 2.71 i 10 2 2.54 i byponovací 2 1.64 i 0.72 i 0.72 i i 1.72 i 1.1 2.1 1.1 2 2.54 i byponovací 2 1.64 i 0.72 i 0.77 i mglczyma 1 2.71 i 0.77 i mglczyma 1 2.71 i 0.77 i 0.77 i mglczyma 1 2.71 i 0.77 i mglczyma 1 1 1 1 1 0.77 i mglczyma 1 1 1 1 1 0.77 i mglczyma 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 <td< td=""><td>audi 10 7.79 minitat 8 4.54 cremely 11 5.64 dyponovać 21 6.44 filat 11 7.77 myjazd 2 5.60 5.03 5.64 dom 10 2.24 renaulit 9 7.77 myjazd 5 12 5.79 gojechać 6 6.14 renaulit 9 7.77 możczyzna 6 1.73 13 2 2.53 6 64.65 ratowicz 7.70 możczyzna 6 1.73 13 2.53 6 64.65 star 9 7.78 możczyzna 6 1.74 6 6 6.44 star 9 7.78 możczyzna 8 0.71 14 2 2.51 5 9 7.64 9 9 BMW 7.76 możcz 8 0.71 30 2 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9</td></td<> <td>🔲 marek</td> <td>1</td> <td>8 7.</td> <td>14</td> <td>transport</td> <td>10</td> <td>4.58</td> <td>63</td> <td>tyle</td> <td>8</td> <td>5.77</td> <td>0</td> <td>poruszać</td> <td>32</td> <td>7.03</td> <td>0</td> <td>kterowca</td> <td>6</td> <td>4.41</td> <td></td> <td></td>	audi 10 7.79 minitat 8 4.54 cremely 11 5.64 dyponovać 21 6.44 filat 11 7.77 myjazd 2 5.60 5.03 5.64 dom 10 2.24 renaulit 9 7.77 myjazd 5 12 5.79 gojechać 6 6.14 renaulit 9 7.77 możczyzna 6 1.73 13 2 2.53 6 64.65 ratowicz 7.70 możczyzna 6 1.73 13 2.53 6 64.65 star 9 7.78 możczyzna 6 1.74 6 6 6.44 star 9 7.78 możczyzna 8 0.71 14 2 2.51 5 9 7.64 9 9 BMW 7.76 możcz 8 0.71 30 2 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	🔲 marek	1	8 7.	14	transport	10	4.58	63	tyle	8	5.77	0	poruszać	32	7.03	0	kterowca	6	4.41		
Inits 12 7.75 Implementation 13 2 5.60 Implementation 1.60 1.2.4 Implementation 7.77 Implementation 1.77 Implementation 1.78 Implementation Implementation 1.78 Implementation	first 12 7.77 wyjszd Z 4.11 wylecy 22 5.00 berować 5.13 32 3.13 32 3.13 32 3.13 32 3.13 3 3.23 3.31<	ibue 🗉	1	0 7.	19	minuta	8	4.54		cztery	13	5.68	1	dysponować	21	6.94	0	mieszkanie	6	3.33		
Pord 8 7.73 0 mcczyma 6 3.73 1 15 9 9.5.11 0 00data/c 6 6.61 ratowniczo 2 7.70 0 doga 11 2.71 5 5.23 0 0data/c 6 4.61 star 9 7.73 0 asoba 11 1.75 1 15 2.21 0 0data/c 6 4.61 star 9 7.73 0 asoba 11 1.75 1 12 2.21 0 0data/c 6 4.61 BMW 8 7.73 0 asoba 11 1.75 2 12 3.16 0 </td <td>Pord 8 7.73 mgkrzyma 6 1.73 1 19 2 3.11 pojechać 6 6.14 Peresult 9 7.72 masto 2 1.1 2.73 0 dotać 6 4.41 star 2 7.73 masto 2 1.70 4 1 2.71 0 dotać 6 4.43 star 2 7.78 masto 2 1.70 4 1 5.12 2.39 dotać 6 4.43 pojechać 6 4.43 pojechać 6 4.43 pojechać 5 7.63 pojechać 6 7.64 pojechać 6 7.64 pojekać 5 7.63 pojekać 5 7.64 pojekać 5</td> <td>💷 fiat</td> <td>3</td> <td>3 7.</td> <td>5</td> <td>wyjazd</td> <td>Z</td> <td>4.11</td> <td>63</td> <td>więcej</td> <td>22</td> <td>5.60</td> <td>10</td> <td>klerować</td> <td>34</td> <td>6.35</td> <td>0</td> <td>dom</td> <td>10</td> <td>2.24</td> <td></td> <td></td>	Pord 8 7.73 mgkrzyma 6 1.73 1 19 2 3.11 pojechać 6 6.14 Peresult 9 7.72 masto 2 1.1 2.73 0 dotać 6 4.41 star 2 7.73 masto 2 1.70 4 1 2.71 0 dotać 6 4.43 star 2 7.78 masto 2 1.70 4 1 5.12 2.39 dotać 6 4.43 pojechać 6 4.43 pojechać 6 4.43 pojechać 5 7.63 pojechać 6 7.64 pojechać 6 7.64 pojekać 5 7.63 pojekać 5 7.64 pojekać 5	💷 fiat	3	3 7.	5	wyjazd	Z	4.11	63	więcej	22	5.60	10	klerować	34	6.35	0	dom	10	2.24		
enalit 9 7.72 drogs 11 2.71 5 12 3.79 drodsh 4.63 enalwit 2 7.78 ministo 1.70 4 12 2.71 0.0012e 6 4.64 star 2 7.78 modes 1.1 1.45 1.14 0.0012e 2.31 0.0012e 2.3.70 0.0012e 0.012 0.0012e 2.3.70 0.0012e 2.3.70 0.0012e 0.012 0.012 0.0012e 2.3.70 0.0012e 0.012 0.0012e 0.012 0.012 0.012 0.012 0.012 0.012 0.012 0.012 0.012 0.012 <td>enailit 9 7.72 droga 11 2.71 5 12 3.79 dotad 6 4.43 ratowniczo 2 7.70 misto 1 1.43 12 2.71 6 6 4.43 BMW 8 7.75 osoba 11 1.44 12 2.71 5 12 5.79 dotracé 6 4.63 BMW 8 7.75 osoba 11 1.44 12 2.11 5.14 50 2 13 5.14 50 2 13 5.14 50 2 13 5.14 50 2 13 5.14 50 2 11 5.14 50 2 5.07 with a 3 3.14 0 2 5.13 50 2 5.07 50 2.04 Polica 10 Polica 10 2.20 0 0 5.04 10 1.43 14.82 14.83 14.176 Polica 12.02 12.04 14.14 14.14 14.14 14.14 14.14 <t< td=""><td>Ford</td><td></td><td>\$ 7.</td><td>3</td><td>mężczyzna</td><td>2</td><td>3.73</td><td>8</td><td>15</td><td>2</td><td>5.51</td><td>13</td><td>pojechać</td><td>2</td><td>6.14</td><td></td><td></td><td></td><td>>></td><td></td><td></td></t<></td>	enailit 9 7.72 droga 11 2.71 5 12 3.79 dotad 6 4.43 ratowniczo 2 7.70 misto 1 1.43 12 2.71 6 6 4.43 BMW 8 7.75 osoba 11 1.44 12 2.71 5 12 5.79 dotracé 6 4.63 BMW 8 7.75 osoba 11 1.44 12 2.11 5.14 50 2 13 5.14 50 2 13 5.14 50 2 13 5.14 50 2 13 5.14 50 2 11 5.14 50 2 5.07 with a 3 3.14 0 2 5.13 50 2 5.07 50 2.04 Polica 10 Polica 10 2.20 0 0 5.04 10 1.43 14.82 14.83 14.176 Polica 12.02 12.04 14.14 14.14 14.14 14.14 14.14 <t< td=""><td>Ford</td><td></td><td>\$ 7.</td><td>3</td><td>mężczyzna</td><td>2</td><td>3.73</td><td>8</td><td>15</td><td>2</td><td>5.51</td><td>13</td><td>pojechać</td><td>2</td><td>6.14</td><td></td><td></td><td></td><td>>></td><td></td><td></td></t<>	Ford		\$ 7.	3	mężczyzna	2	3.73	8	15	2	5.51	13	pojechać	2	6.14				>>		
instamiczo Z 7.70 instato Z 1.70 4 12 2.51 instato 2.770 instato Z 1.70 4 12 2.51 instato 2.770 instato Z 1.70 4 12 2.51 instato 2.770 instato Z 1.70 1 4 12 5.21 0/07 0/0	Instanticize 2 7.70 mistato 2 1.70 4 12 2.51 0 dotrace 6 4.41 Immediate 1.4.45 Immediate Immediate 1.4.45 Immediate <	🗐 renault		9 7.	2 0	droga	11	2.71	8	5	17	5.29	0	dostač	6	4.65			-	_		
star 2 7.58 0x00a 12 1.14 0x0 5.70 BMW 8 7.36 0x0a 8 0.71 2 12 3.16 0x0a 0x1an 6 7.36 0x0a 8 0.71 2 3.05 2 3.16 0x1an 6 7.05 0x0a 2 5.00 2 5.11 0x0a 0x0a 2 3.07 0x1an 6 7.05 0x0a 2 5.00 2 5.11 0x0a 0x0a 2 3.01 0x1an 6 7.05 0x0a 2 5.00 2 5.01 0x0a 0x0a 2 5.00 0x0a 0x0a <t< td=""><td>star 2 7.58 oxoba 1.1 1.48 6 6 5.11 5.74 poble 2 2.70 BMW 8 7.7.56 praca 8 0.71 5 5 7.11 5</td><td>ratowniczo</td><td></td><td>Ζ 7.</td><td>no 💿</td><td>miasto</td><td>Z</td><td>1.70</td><td>63</td><td>4</td><td>12</td><td>5.21</td><td>8</td><td>dotrzeć</td><td>6</td><td>4.43</td><td>EX.</td><td>ecz_nad-w-l-r</td><td>m 34</td><td>1.30</td><td></td><td></td></t<>	star 2 7.58 oxoba 1.1 1.48 6 6 5.11 5.74 poble 2 2.70 BMW 8 7.7.56 praca 8 0.71 5 5 7.11 5	ratowniczo		Ζ 7.	no 💿	miasto	Z	1.70	63	4	12	5.21	8	dotrzeć	6	4.43	EX.	ecz_nad-w-l-r	m 34	1.30		
BMW 8 7.36 Prsca 8 0.71 2 12 3.14 2004dc 9 0.43 intisan 6 7.36 50 2 5.61 >>> >>> wersja 5 1.14 merceles 6 7.75 2 9 5.11 duio 7 5.04 >>> wersja 5 1.14 ceracles 6 7.75 2 9 5.01 >> 1.42 volkswagen 6 7.60 Posta 19 2.02 6 7.60 1.14 1.42 volkswagen 6 7.60 Posta 19 2.02 6 6.61 0.64 1.42 volkswagen 6 7.60 Posta 19 1.81 Wordlaw 10 6 4.62 przybadek 2 3.07 1.41 1.76 Wordlaw 19 4.42 3 8 4.05 przybadek 1.20 5	BMW 8 7.36 praca 8 0.71 2 12 2.18 200162 9 0.43 ktasa 8 3.27 Initisan 6 7.36 30 7 5.66 3.04 email email 2.2.50 Barcelona 2 7.16 30 7 5.66 email 12 2.02 Gdariok 22 6.61 6 6 7.464 integrade 7 6 1.48 Workzawa 10 4.96 30 7 4.64 integrade 7 6.18 integrade 5.18 integrade 5.18 integrade 5.18 integrade 5.18 integrade 5.12 integrade 5.18 integrade 5.18 integrade 5.18 integrade 5.18 integrade 5.18 integrade 5.18 integrade integrade 5.18 integrade 5.18 integrade	star		9 7.	18	osoba	11	1.45	6	ile	6	5.21	0	byc	81	1.54	10	pobliże		Z 5.70		
mison 6 7.36 >> 50 7 50 7 50 7 50 7 50 7 50 7 50 7 50 7 50 7 <th7< th=""> <th7< th=""> <th7< th=""></th7<></th7<></th7<>	nisan 6 7.36 >> 30 7 5.18 >>> wersja 8 3.14 mercedes 6 7.00 20 2.513	BMW		\$ 7.	16	praca	8	0.71		2	12	5.18	6	zostać	\$	0.43	0	klasa		8 3.27		
Barcelona 2 7.18 20 2 5.13 Icerational 2 2.00 Odarisk 22 6.66 6 6 7.00 Icerational 7.60 Poliska 10 20 2 5.13 Icerational 7.60 Poliska 10 20 2 5.13 Icerational Icerational 7.60 Poliska 10 20 2 5.14 Icerational Icerational 7.60 Poliska 10 2.00 Poliska 10 10 6 4.60 Icerational Icerational 11.85 Icerational	Barestona 2 7.18 20 2 5.10 cena 7 2.30 Godshk 22 5.66 6 6.62 Polska 10 2.02 Godshk 22 5.66 6 6.62 Polska 10 2.02 Workswagen 6 7.40 Polska 10 2.02 Workswagen 100 6 4.42 co 7.46 Polska 10 1.81 Workswagen 10 4.40 kategoria 7 2.46 meijsce 12 1.85 Workswagen 10 4.40 przypadci 2.06 przypadci 2.06 Vorkswagen 10 4.42 9 4.05 przypadci 2.06 >>> Ito 15 0.31 wiele 28 4.05 polska 6 1.28 >>> Lo	U nissan		6 7.	16			>>	10	50	Z	5.18				>>	0	wersja		1 3.14		
mercedes \$ 7.05 0.020 \$ 2.06 Teter, 1.06 \$ 7.05 0.020 \$ 2.02 Gdatida \$ 22 6.66 Ikika \$ 11 \$ 4.20 \$ 7.05 Ikika \$ 12.02 Gdatida \$ 22 6.66 Ikika \$ 11 \$ 4.20 \$ 7.05 Ikika \$ 12.02 Gdatida \$ 22 6.66 Ikika \$ 11 \$ 4.20 Ikika \$ 1.422 Ikika \$ 1.432 Ikika Ikika \$ 1.432 Ikika Ikika \$ 1.452 Ikika \$ 1.452 Ikika Ikika \$ 1.452 Ikika	mercedes 9 7.00 0.00 7.00 recer_point recer_point recer_point recer_point recer_point recer_point recer point recer	Barcelona		Z 7.	8				10	20	2	5.13	5				0	cena		Z 2.50		
U Odotki, 42 6.0-00 0.004 (2) 9.0-50 0.004 (2) 9.0-50 0.004 (2) 9.0-50 Rzerzów 2 6.11 100 6 4.80 0.4400 (2) 4.450 (2	0 datak 24 0.04 21 0.04 100 6 1.85 Farestow 2 0.05 0.00 0.00 0.00 1.00 6 1.00 6 1.00 6 1.00	mercedes		9 7.	10				0.0	duzo	1	5.04	123	ez pod rzecz n	n 47.	0.60		Polska		10 2.02		
Autow 2 0.11 10 0 0.00 kategoria 2 3.48 mejsce 12 1.83 Wanzawa 23 5.26 30 2 4.44 0 co 2 3.01 fmas 13 1.76 Wroclaw 10 4.96 6 6 6.462 przypadek 2 2.06 >> Krahów 10 4.42 3 8 4.06 państwo 6 1.40 >> to 15 0.31 wiele 28 4.05 Polaka 6 1.28 >>> >> >> >> >> >> >>	Arktow 2 0.11 100 0 -00 istategrin 2 3.48 miejsce 12 1.45 Workzaw 10 3.0 7 4.64 istategrin 2 2.48 miejsce 12 1.45 Workzaw 10 4.96 6 5 4.62 przybadz 2.06 >>> to 13 8 4.66 połstwo 6 1.28 to 13 0.31 wiele 2.8 4.05 połstwo 6 1.28 io 12 0.60 io 12 0.60 io io 1.28 viele 2.8 4.05 połstwo 6 1.28 io io io	Gdansk		0.	0				10	KIUKA	11	4.82	1	volkswagen	3	6 7.40	10	kraj		6 1.88		
matana 22 3.07 1 mma 13 1.78 Wroclaw 10 4.96 6 6 4.22 przypadek 2 3.07 Kraków 10 4.42 przypadek 2 3.07 1 mma 13 1.78 Kraków 10 4.42 3 8 4.06 patrixwo 6 1.40 to 15 0.31 wiele 2.8 4.05 1.28 >> >> 12 0.65	Windtaw 10 4.50 0 1 vvi co 2 3.07 firma 14 1.76 Windtaw 10 4.96 0 5 4.62 0 1.40 >>> >>> Krahów 10 4.42 3 8 4.66 patritivo 6 1.40 >>> to 15 0.31 wiele 28 4.05 ccas 6 0.23 >> >>> vert pod cractowium 33 1.00 ract pod cractowium 12 1.00 ract pod cractowium 1.00 ract pod cractowium 12 1.00 ract pod cractowium 12 1.00 ract pod cractowium 1.00 ract pod cractowium 1.00 <td>Warrawa</td> <td></td> <td>1 0.</td> <td>4</td> <td></td> <td></td> <td></td> <td>0</td> <td>100</td> <td>0</td> <td>4.64</td> <td>12</td> <td>kategoria</td> <td>2</td> <td>Z 3.48</td> <td>12</td> <td>miejsce</td> <td></td> <td>12 1.85</td> <td></td> <td></td>	Warrawa		1 0.	4				0	100	0	4.64	12	kategoria	2	Z 3.48	12	miejsce		12 1.85		
Workshim 10 4.42 0 3 8 4.06 1.40 to 15 0.31 0.41 0 9 1.40 >>> >> 0 10 1.42 0 1.40 to 15 0.31 0 0 12 0.65 >>> >>> >>> >>> >>>	mickam ji s to prppadk ji ji <td>Winstawa</td> <td></td> <td>0 4</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>6</td> <td>4</td> <td>4.67</td> <td></td> <td>0</td> <td>-</td> <td>Z 3.07</td> <td>11</td> <td>firma</td> <td></td> <td>1.76</td> <td></td> <td></td>	Winstawa		0 4						6	4	4.67		0	-	Z 3.07	11	firma		1.76		
to 15 0.31 to 15 0.31 >>> wiele 28 4.05 Polaka § 1.28 >> >>>	wiele 28 4.05 5 1.40 to 15 0.31 wiele 28 4.05 6 1.28 >>> to 19 0.45 6 0.33 0.33 0.33 rear pod creationium 332 100 reart pod creationium 176 100 100 100 100 100 100 100 100 100 100 100 100 <t< td=""><td>Kraków</td><td></td><td>0 4</td><td>0</td><td></td><td></td><td></td><td>0</td><td>1</td><td></td><td>4.04</td><td></td><td>przypadek</td><td></td><td>1 2.06</td><td></td><td></td><td></td><td>>></td><td></td><td></td></t<>	Kraków		0 4	0				0	1		4.04		przypadek		1 2.06				>>		
>> Polkad \$ 1.7.8 >> to 12 0.65 C2a5 \$ 0.7.3	>> >><	to		5 0	1				100	wiele	28	4.05		panstwo	8	9 1.40						
		1.00	1	-					122	100725			10	Polska		2 1.28						
				1.00									10	(74)	1	6 0.03						
<u>>></u>													100			0.23						
	Terr pod marnavims 117 1 10 mar pod mar h 740 170 mar pod marnavih 199 2 30 mar padnavims 176 3.00 mar pod manualih 99 1 30															>>						
	Terr pod cractowiems 112 1 10 [cract pod cract b 740 1 70] [cract pod cractonaldh 199 2 20] [cract pod cractonaldh 199 1 20]												-									
goona 8 5.99 dvpozycla 7 5.15 eara2 25 8.65 parkine 23 7.75 upywad 8 3.70				- 2		1	10 C		-	100-2				4 7 04				11 4 10			 Lash I	





3. Links

- Sketch Engine: <u>http://www.sketchengine.co.uk/</u>
- Slovene sketch grammar: <u>https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf</u>
- Narodowy korpus języka polskiego (National Corpus of Polish): <u>http://nkjp.pl/index.php</u>
- Wielki słownik języka polskiego (Great Dictionary of the Polish Language): <u>http://www.wsjp.pl/</u>
- Instytut języka polskiego (Polish Language Institute): <u>https://www.ijp-pan.krakow.pl/</u>
- Zespół Inżynierii Lingwistycznej (Linguistic Engineering Group): <u>http://zil.ipipan.waw.pl/</u>
- Słowosieć (plWordnet): <u>http://plwordnet.pwr.wroc.pl/wordnet</u>