Mariusz Piotrowski Węzeł Centralny OŻK-SB

Analiza frekwencyjna słów używanych w dokumentach JST na przykładzie "Strategia marki i promocji miasta Augustowa w latach 2010-2015". *Instrukcja użytkowania programu Antconc dla korpusów tekstowych języka polskiego*.

Warszawa, 2014

1. Wprowadzenie.

Program do analiz językowych można pobrać ze strony twórcy, pod adresem <u>http://www.laurenceanthony.net/</u> <u>software.html</u>

Działa on na najpopularniejszych systemach operacyjnych, **Microsoft Windows**, **Mac OSX** i **Linux**. Instrukcja bazuje na wersji programu 3.4.3.

2. Przygotowanie środowiska pracy.

Korzystanie z programu **Antconc** należy rozpocząć od ustawienia parametrów środowiska pracy. Pierwszą czynnością powinno być ustawienie sposobu kodowania polskich znaków.





głównym przechodzimy do opcji Settings, następnie Global Settings.

Aby móc pracować na polskich korpusach tekstów należy zmienić w kategorii **Character Encoding - Current Encoding -Unicode (UTF-8).** (Taki sposób kodowania jest domyślny na większości współczesnych systemów operacyjnych). Jeśli w okienku **Current Encoding** jest kodowanie inne (np. CP-1250) wówczas należy wcisnąć **Edit** i wybrać **UTF-8.** Zmiany zatwierdzamy wciskając **Apply**.

3. Przygotowanie plików do analizy.

Program **Antoconc** pozwala na analizy korpusów tekstowych zapisanych jako proste pliki tekstowe. Pakiet **Microsoft Office** w trakcie wykonywanych czynności miał problem z eksportowaniem plików do formatu tekstowego. Dlatego też polecam wykorzystać w tym celu program Writer znajdujący się pakiecie <u>Libreoffice.</u> Program jest darmowy i licencjonowany na jednej z wolnych licencji.

Dokumenty strategiczne, które są udostępniane w formatach do edycji- np.: **docx**, czy **odt** należy po otwarciu zapisać jako dokument tekstowy. Natomiast dokumenty, które są dystrybuowane jako pliki **pdf**, należy poddać obróbce programem typu **OCR**, który rozpozna zawartą treść i pozwoli dokument zapisać jako plik do edycji.

Tak przygotowany plik z korpusem tekstu możemy załadować do naszego programu. W tym celu wybieramy **File** i **Open File(s)**.

🗯 AntConc	File Settings Help						- 🕹	● 0 □	😺 🤶 🕬	99% [/]• p	on. 12:31	Q :≡
	Open File(s)	^F	AntConc 3.4.3r	n (Macintosh	OS X) 2014							
Corpus Files	Open Dir	^ D Concor	dance Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List				
	Close Selected File(s) Close All Files								,		File	
	Clear Tool Clear All Tools Clear All Tools and Files											
	Save Output to Text File	^S										
'	Import Settings from File Export Settings To File											
	Restore Default Settings											
	Exit											
	Search Term 🗸 Words	Case Regex	Search Window Size									
		Advanced	50									
Total No.	Start Stop Kwic Sort	Sort										
0 Files Processed	Level 1 1R	Level 2 2R 🗘 🗸 Level 3 3R	0								Clone Re	sults
	-											

Pojawia się manager plików. Wybieramy stworzony przez nas plik z korpusem w formacie txt.



Poprawnie załadowany plik powinien pojawić się w **menu bocznym**.



4. Analiza frekwencyjna.

Po wykonaniu powyższych czynności można już przystąpić do pierwszych analiz.

Analiza frekwencyjna, wraz z wyszukiwaniem powtarzalnych połączeń wyrazowych (kolokacji) są podstawowymi technikami badawczymi na gruncie NLP (*Natural Language Processing*) dziedziny korzystającej z językoznawstwa i informatyki.

Pierwszym krokiem jest przejście do zakładki **Word List**. I następnie wciśnięcie przycisku **Start**.

🗯 AntConc File	Settings Help	🛇 🔲 😻 🛜 🕪) 78% 🔳 pon. 15:0	6 Q ☷
	AntConc 3.4.3m (Macintosh OS X) 2014		
Corpus Files	Concertance Concertance Dist File View Churteen & Collectors Ward List Key	ward List	
strategia_augustow.txt	Concordance Plot Prie view Clusters/N-Grams Collocates Word List Key	word List	
	Rank Freq Word Word Word Verse V Search Hits: 0		
	Search Term V Words Case Regex Hit Location		
	Start Stop Sort Lemma List Loaded		
Total No.	Sort by Invert Urger		
1 Files Pressed	Sort by Freq	Clon	e Results
4			

W efekcie otrzymujemy surową listę frekwencyjną słów z korpusu tekstowego. W tekście eliminujemy zaimki, spójniki i przyimki. Z punktu widzenia naszych analiz nie noszą one interesujących treści.



Aby wyczyścić tekst z tego typu niepożądanych słów można skonstruować **Stoplistę**. Jest to zbiór słów, które mają być wyłączone z procesu przetwarzania i nie będą się pojawiać w dalszych analizach. Przykładowa stoplista znajduje się w zasobach wikipedii (<u>http://pl.wikipedia.org/wiki/Wikipedia:Stopwords</u>). Proszę kliknąć w plik z naszą <u>stoplistą</u>. Struktura pliku jest bardzo prosta, słowa są oddzielone przecinkami. Można więc ten plik dowolnie rozbudowywać, bądź też usuwać słowa, jeśli uznamy, że jest on zbyt rozbudowany.

Aby załadować **stoplistę** należy w menu głównym wybrać **Settings** i następnie **Tool Preferences**.

🛎 AntConc File	Settings Help								90	😺 🤶 🜒) 609	6 🔳 🛛 pon. 15:51	ଏ ≔
	Global Settings			AntConc 3.4.3r	n (Macintos	h OS X) 2014						
Corpus Files	Tool Preference	95	Conservation	Casaardaaaa Diat	File Menu	Chusters (N. Orano	Collegates	Ward Liet	Keyward Liet			
strategia_augustow.txt			Concordance	Concordance Plot	File view	Clusters/IN-Grams	Conocates	Word List	Reyword List			
	Bank Freq	Word Tokens: 7520	Search Hits:	0		Lemma Word Form(s	6					
	1 736	w					<i>'</i>					
	2 224	i										
	3 133	na										
	4 100	Z										
	5 91	do										
	6 87	się										
	7 63	augustów										
	8 57	marki										
	9 55	jest										
	10 54	miasta										
	11 49	augustowa										
	12 46	to										
	13 40	effective										
	14 40	launching										
	15 35	oraz										
	16 34	można										
	17 34	nie										
	18 33	0										
	19 32	ara										
	21 26	iako										
	22 26	projekt										
	23 25	po										
	24 24	należy										
	25 24	np										
	Search Term		Hit Locativ									
		Advance	d Search	Only 0								
	Start	Stop Sort	Lemma Li	st Loaded								
Total No.	Sort by Invert	Order										
1	Sort by Freq	~									Clone	Results
Files Processed												

W oknie, które się pojawia wybieramy Word List.



Przechodzimy do fragmentu **Word List Range**. Ustawiamy wartość **Use a stoplist below** i w opcji **Add Words From File** naciskamy **Open**. Wybieramy nasz plik **stoplist.txt**.

Atti Word Types: Rank F 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	3029WordreqWord236w224i133na100z91do87się63augustów57marki55jest54miasta49augustówa46to40effective40launching35oraz	Conco Ti Category Concordance Clusters/N-Grams Collocates Word List Keyword List	AntCone 3.4.3m (Macintosh OS X) 2014
Altor Al	3029WordreqWord236w224i133na100z91do87się63augustów57marki55jest54miasta49augustówa46to40effective40launching35oraz	Tr. Conco Category Concordance Clusters/N-Grams Collocates Word List Keyword List	Incordance Plot File View Clusters/N-Grams Collocates Word List Keyword List Tool Preferences Vord List Preferences Verd List Prequency Verd Lemma Word Form(s) Other Options Vere at all data as lowercase Treat case in sort Lemma List Loaded Load Treat Word List Range as Lemma List Range Word List Range Use specific words below Add Word Add Add Word From File Open
Viol Types: Rank F 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	3029WordrreqWord236w224i133na100z91do87się63augustów57marki55jest54miasta49augustówa46to40effective40launching35oraz	T Category Concordance Clusters/N-Grams Collocates Word List Keyword List	Word List Preferences Display Options
Word Types: Rank F 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	3029WordreqWord236w224i133na100z91do87się63augustów57marki55jest54miasta49augustowa46to40effective40launching35oraz	Category Concordance Clusters/N-Grams Collocates Word List Keyword List	Word List Preferences Display Options
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	236 w 224 i 133 na 100 z 91 do 87 się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	Category Concordance Clusters/N-Grams Collocates Word List Keyword List	Word List Preferences Display Options
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	256 W 224 i 133 na 130 z 91 do 87 się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	Concordance Clusters/N-Grams Collocates Word List Keyword List	Display Options Image: Second stress in sort Lemma List Loaded Load Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Add Word Add Add Word Open
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	224 1 133 na 130 z 91 do 87 się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	Collocates Word List Keyword List	 Hank Vord Frequency Vord Lemma Word Form(s) Other Options Treat all data as lowercase Treat case in sort Lemma List Loaded Load Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Use a stoplist below Add Add Add Word Add
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	133 na 100 z 91 do 87 się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	Word List Keyword List	Other Options Image: Treat all data as lowercase Treat case in sort Lemma List Loaded Loaded Loaded Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Add Word Add Add Word From File Open
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	100291do87się63augustów57marki55jest54miasta49augustowa46to40effective40launching35oraz	Keyword List a g	Treat all data as lowercase Treat case in sort Lemma List Loaded Loaded Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Add Add Word Add Open
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	91 do 87 się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Treat case in sort Lemma List Loaded Load Load Load Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Add Word Add Add Add
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	 87 Się 63 augustów 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz 	a e g	Lemma List Loaded Load Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Add Word Add Add Mord Add Copen
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21	63 augustow 57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Lemma List Loaded Load Treat Word List Range as Lemma List Range Word List Range Use all words Use specific words below Ouse a stoplist below Add Word Add Add Open
8 9 10 11 12 13 14 15 16 17 18 19 20 21	57 marki 55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Coaded Coaded Treat Word List Range Use all words Use specific words below Add Word Add Add Words From File
9 10 11 12 13 14 15 16 17 18 19 20 21	55 jest 54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Treat Word List Range Word List Range Use all words Use specific words below Add Word Add Add Words From File Open
10 11 12 13 14 15 16 17 18 19 20 21	54 miasta 49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Word List Range Use all words Use specific words below Add Word Add Add Words From File Open
11 12 13 14 15 16 17 18 19 20 21	49 augustowa 46 to 40 effective 40 launching 35 oraz	a e g	Use all words Use specific words below Add Word Add Add Words From File Open
12 13 14 15 16 17 18 19 20 21	46 to 40 effective 40 launching 35 oraz	e 9	Add Word Add Add Add Words From File Open
13 14 15 16 17 18 19 20 21	40 effective 40 launching 35 oraz	e g	Add Words From File Open
14 15 16 17 18 19 20 21	40 launching 35 oraz	9	Add Words From File Open
15 16 17 18 19 20 21	35 oraz		
16 17 18 19 20 21			
17 18 19 20 21	34 można		Clear
18 19 20 21	34 nie		
19 20 21	33 o		
20 21	32 dla		
21	27 emo		
	26 jako		
22	26 projekt		
23	25 po		
24	24 należy		
25	24 np		
Search Term	Words Case	3	Apply Cancel
		Advanced	Search Only 0 0
Start	Stop	Sort Le	Lemma List Loaded
Sort by	Invert Order		
Sort by Free			Clone Res

Jeśli czynność zostanie zakończona sukcesem słowa ze stoplisty powinny się załadować do programu.

Wybór potwierdzamy przyciskiem **Apply**.



Należy następnie jeszcze raz przeładować listę, dopiero wówczas pojawią się frekwencje słów, bez słów ze stoplisty.

单 🛋 🛋	onc Fil	e S	Settings	Help								😻 🗩 🕤	🗋 🤶 🜒) 46	6% 🔳 🕐 pon. 16:	18 Q ≔
							AntConc 3.4.3r	n (Macintos	h OS X) 2014						
Corpus Files						Conservations	Orange Dist	File Manu	Churcherry (h) Commo	Orlinenter	Manual Lint	Kennedlist			
strategia_august	ow.txt					Concordance	Concordance Plot	File view	Clusters/IN-Grams	Collocates	word List	Keyword List			
			Word Type Bank	es: 2873	Word Tokens: 5515	Search Hits:	0		Lemma Word Form(s	a)					
		1	1	62	augustán.				Lennia Word Forma	9					
			2	57	manki										
			2	54	murkt										
			4	49	auaustowa										
			5	40	effective										
			6	40	launchina										
			7	27	emo										
			8	26	projekt										
			9	24	należy										
			10	24	np										
			11	20	miasto										
			12	19	projektu										
			13	17	działań										
			14	17	promocji										
			15	16	augustowie										
			16	16	produkt										
			17	15	cele										
			18	15	pozycjonowanie										
			19	15	reklamy										
			20	15	trzy										
			21	14	augustowski										
			22	14	big										
			23	14	miejsca										
			24	14	poprzez										
			25	14	strategii										
			Search Te	erm 🔽 Wo	ords 🗌 Case 📄 Regex	Hit Locatio	on								
					Advanc	ed Search	Only 0								
			Start		Stop Sort	Lemma Lis	st Loaded								
Total No.			Sort by	Invert O	Irder										
1			Sort by F	req	×									Clo	one Results
Files Processed															

Praca ze słownikiem języka polskiego

Tak wyodrębnione frekwencje nie uwzględniają jednak różnorodnych końcówek fleksyjnych. Jak widać na wcześniejszym obrazie - 63 razy pojawia się Augustów, 49 - Augustowa, 16 - Augostowie, 14 - augustowski. W celu właściwego określenia frekwencji korpusu tekstowego należy zliczyć więc formy podstawowe słów, w językoznawstwie zwane - *lemma*. Osobno zostaną zliczone rzeczowniki, czasowniki i przymiotniki. Język polski jest językiem fleksyjnym, z dużą liczbą wyjątków, co sprawia, że nie można zastosować procesu *stemmingu*, czyli programów tworzących formy fleksyjne w oparciu o ustalone reguły.

W celu obejścia tego problemu można posłużyć się przygotowanym słownikiem języka polskiego. Słownik ten został przygotowany w oparciu o zasoby **aspell**. Słownik znajduje się w <u>zakładce teksty ożk</u>, wraz z plikiem stoplista.txt. Składa się on dotąd ze 194 tys. form podstawowych, i ponad 3,6 mln słów w różnych formach fleksyjnych.

Uwaga: ze względu na rozmiar pliku ze słownikiem- ponad 55 Mb, do pracy rekomendowany jest komputer posiadający co najmniej 4 Gb pamięci operacyjnej (RAM).

Aby załadować słownik języka polskiego do programu **Antconc** należy przejść do menu głównego i wybrać **Settings** i **Tool Preferences**, (por. strona 13) i przejść do opcji **Word List** (por. strona 14). W obszarze **Lemma List** wciskamy **Load** i

	AntConc 3.4.3m (Macintosh OS X) 2014
ous Files	Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
tegia_augustow.txt	
	word types: 26/3 word lokens: 35/5 Search Hits: 0 Rank Free Word I emma Word Form(s)
	2 57 model
	2 J7 Multiku 3 54 milosto
	+ + + ugus cova looi Preterences
	5 40 effective Category Word List Preferences
	7 37 oros Concordance Display Options
	7 27 emb Clusters/N-Grams ✓ Rank ✓ Frequency ✓ Word ✓ Lemma Word Form(s)
	o zo projekt Collocates Other Options
	10 24 north Kanword List Virat all data as lowercase
	10 24 np riste
	11 20 militor
	12 13 projektu Lemma List
	13 17 aztatan
	15 16 ground and a standard and a standa
	15 16 adjustovice Word List Bange
	10 10 produkt Use all words Use specific words below Use a stoplist below
	17 15 Cele
	Add Words From File stoplista.txt Open
	20 15 trzy
	21 14 augustowski aby Clear
	22 14 blg ach
	23 14 mrejsca aczkolwiek
	24 14 poprzez
	25 14 strategil
	Search Term V Words Case 1
	Start Stop Sort
10.	Sort by Invert Order
	Sort by Freq V Cancel Clone Results
rocessed	

wczytujemy nasz słownik języka polskiego. Proces ten jest czasochłonny, i może trwać kilka minut, w zależności od szybkości komputera. W tym czasie program może nie odpowiadać.

🔹 AntConc F	ile Sett	ings He	elp				🐯 🥯 🛇	📋 🤶 🕪)) 86% [⁄-]) pon. 18:	58 ् ः≡
0 🔴					AntConc	3.4.3m (Macintosh OS X) 2014			
Corpus Files					-				
strategia_augustow.txt					in:	strukcja antconc	Q Search	J	
	W	ord Types:	2873 Word	Fauncitan		wrzesień			
	1	C 2		Favorites		Analiza frekwyka polskiego	3D,->		
	1	. 63	s aug	u 👽 Dropbox	onc	lipiec	4chanami, 4chanem, 4chanie,		
	2	. 57	nar 1 mia	🖕 🔲 Wszystkie moje pliki	17 o 10.43.50	Strategia mar010-2015.pdf	4chanom, 4chanowi, 4chanów,		
		, J4	+ III.u		17 o 10.44.51	2013	4chany a>		
	5	· · · · · ·	aug a off		17 o 12.31.55	h odm_polska.txt.zip	a.,->		
	6	40	a 1au	 y→y Programy 	17 o 12.32.00	2012	A,->		
	7	27	7 emo	Development	17 o 15.06.30	odm_polska.txt	a capite,->		
	8	26	5 pro	i 🔜 Biurko	17 o 15.07.22	etoplista tyt	a cappella,->		
	9	24	4 nal		17 o 15.51.50	Stophota.txt	a contrario,->		
	1	.0 24	4 np		17 0 15.52.05		a cunabulis,->		
	1	.1 20	0 mia	s 🛅 Dokumenty	17 0 15.52.23		a discretion,-> a fortiori>		
	1	.2 19	9 pro	j 💽 Downloads	17 0 15.53.41		a fresco,->		
	1	.3 17	7 dzi	a	17 0 16 19 07		a fronte,->		
	1	.4 17	7 pro	m			odm polska.txt		
	1	.5 16	5 aug	u					
	1	.6 16	5 pro	d			text - 56,7 MB		
	1	.7 15	5 cel	e			Created 2 lis 2012, 22:30		
	1	.8 15	5 poz	У			Modified 2 lis 2012, 22:30		
	1	.9 15	5 rek	1			Last opened Dzisiaj, 18:48		
	2	0 15	5 trz	У			Add Tags		
	2	1 14	4 aug	u					
	2	2 14	t big						
	2	.5 14 14 14	+ mie]			Caraal		
	2	.4 14	+ μομ 1. s+n				Caricei		
	-								
	Se	earch Term	V Words	Case Regex Hit L	ocation				
				Advanced S	earch Only 0	0			
		Start	Stop	Sort Lem	ma List Loaded	1			
Total No	So	ort by 📃 I	Invert Order						
1	S	ort by Freq	`	·				Clo	ne Results
Files Processed	-								

Jeśli operacja przebiegnie pomyślnie ukaże się ekran, z formami fleksyjnymi. Teraz wystarczy nacisnąć **OK**, a następnie

🗯 AntConc File	Settings Help								V 🔊 🖓 🗸) 🤶 🌒 8	7% [4]• pon. 19	9:02 ् ः≡
000				AntConc 3.4.3r	n (Macintos	h OS X) 2014						
Corpus Files							-					
strategia_augustow.txt			Concordance	Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List			
	Word Types: 2873	Word Tokens: 5515	Search Hits:	0		Lamma Ward Farm/	-)					
	Hank Freq	word				Lemma word Porm(s	5)					
	1 63	augustow										
	2 57	marki										
	3 54				Lemm	a List Entries: 363	31427					
	4 49 5 40				Lonnin						_	
	6 40	lounching a->										
	7 27	emo A->										
	8 26	projekt a battuta-	>									
	9 24	należy a capite->										
	10 24	np a cappella	->									
	11 20	miasto a contrari	0->									
	12 19	projektu a cunabuli	s->									
	13 17	działań a discréti	on->									
	14 17	promocji a fresco-	->									
	15 16	augustowie a fronte->										
	16 16	produkt a giorno->										
	17 15	cele a kuku->										
	18 15	pozycjonowar a limine->										
	19 15	reklamy a linea->										
	20 15	trzy a maiore a	d minus->									
	21 14	augustowski a nużby->										
	22 14	big a nuże->										
	23 14	miejsca a piacere	>.									
	24 14	poprzez a posterio	r1->									
	25 14	strategii										
	~							OK				
	Search Term V	Vords Case Stop Sort							~			
Total No.	Sort by Invert	Order										
1 Files Processed	Sort by Freq	~										Clone Results

nacisnąć **Apply**.

Pozostaje jeszcze ponownie wygenerować listę słów - przyciskiem **Start**. Efekt jest następujący.

🗯 AntConc File	Settings Help				🐯 🧼 🛇 🔲 🛜 🕪)) 88% [分) pon. 19:03 🔍 😑
			AntConc 3.4.3r	n (Macintos	sh OS X) 2014
Corpus Files			Concerdance Dist	File Manu	Okustan (d. Osama – Osllanstan – Ward List – Kasunad List
strategia_augustow.txt			Concordance Plot	File View	Clusters/N-Grams Collocates Word List Reyword List
	Word Types: 1947 Bank Free	Word Tokens: 5515	Search Hits: 0	0	Lemma Word Form(s)
	1 120	augustáw		(
	2 61	miasto-naństwo			migst 5 migsta 54 migstami 2
	3 57	markowie			merki 57
	4 57	projekt			projekt 26 projektach 1 projektu 19 projekty 2 projektów 9
	5 43	wczasowo-turystyczny			turvstyczna 3 turvstyczne 12 turvstycznego 4 turvstycznej 10 turvstycznych 11 turv
	6 42	produkt			produkcie 3 produkt 16 produktami 1 produktu 8 produkty 5 produktów 9
	7 40	effective			to any other and the angle of t
	8 40	launching			
	9 35	augustowski			augustowska 7 augustowski 14 augustowskich 4 augustowskiej 6 augustowskim 3 august
	10 30	pozycjonowanie			pozycjonowania 13 pozycjonowanie 15 pozycjonowaniem 2
	11 30	promocyjny			promocyjna 1 promocyjne 9 promocyjnego 3 promocyjnej 1 promocyjny 2 promocyjnych 1
	12 29	komunikacja			komunikacja 10 komunikacji 11 komunikacją 1 komunikację 7
	13 27	emo			emo 27
	14 26	działanie			działania 7 działaniami 1 działaniem 1 działań 17
	15 25	budować			budowania 2 budowanie 12 budować 5 budowała 1 buduje 4 budują 1
	16 25	należy			należałoby 1 należy 24
	17 25	rekomendować			rekomendowana 1 rekomendowane 5 rekomendowanego 1 rekomendowanych 1 rekomenduje 4
	18 24	impreza			imprez 5 impreza 9 imprezach 2 imprezami 1 imprezy 7
	19 24	mlejsce			miejsc 2 miejsca 14 miejscach 1 miejsce 3 miejscem 3 miejscom 1
	20 24	np			np 24
	21 24	turystyka			turystyka 14 turystyki 6 turystyką 2 turystykę 2
	22 23	promocja			promocju z promocji i promocjų i promocjų z promocyj i
	23 21	kluczowy			kluczowa 12 kluczowach 5 kluczowam 4
	25 21	reklany			reklam 6 reklamv 15
	22 21	· · ·			
	Search Term 🔽 W	/ords Case Regex	Hit Location		
		Advance	ed Search Only 0 🗘		
	Start	Stop Sort	Lemma List 🗸 Loaded		
Total No.	Sort by Invert	Order			
1	Sort by Freq	~			Cione Results
Files Processed					

Dzięki zastosowaniu słownika języka polskiego oraz stopliście udało się zredukować listę słów o 1000 pozycji. Wynosi ona teraz 1947 pozycji (na początku było to 3029 słów). Jednak całego procesu nie można zautomatyzować. Pojawiają się błędy, które wynikają z samej natury automatyzacji procesów analizy językowej.

Częstym typem błędu jest pojawienie się wyrazów połączonych myślnikiem, mimo tego, że w samym tekście takie zwroty nie funkcjonowały. W analizowanym korpusie było to *miasto-państwo* i *wczasowo-turystyczny*. Natomiast dość łatwo można dojść do źródeł tych słów - podgląd form, które zsumowały się na liczebność formy podstawowej można zobaczyć w głównym oknie w kolumnie **Lemma Word Form(s)**. I można przekonać się, że *miasto-państwo* to faktycznie rożne formy rzeczownika *miasto, a wczasowo-turystyczny* to różne formy przymiotnika *-turystyczny*.

Natura automatyzacji procesu analizy językowej sprawia, że pojawiają się nieprawidłowo dobrane ze względu na funkcję słowa formy podstawowe. Na 3 miejscu pojawiło się słowo *markowie* - jako lemma słowa *marki* (z kontekstu tekstu wynika, że jest to wyraz *marka* - jako znak rozpoznawczy produktu w liczbie mnogiej). Jest to aspekt, który wymaga od badacza dużej czujności w trakcie wyciągania wniosków o naturze tekstu.

Wreszcie mogą pojawiać się niezrozumiałe w pierwszym momencie ciągi znaków. Program Antconc pozwala na szybki podgląd słowa, i określenie jego kontekstu.

Słowo, które chcemy zbadać można kliknąć. Wystarczy w tym celu wybrać słowo w kolumnie Lemma. Np. *emo*.

🗯 AntConc Fi	le Settings He	lp		😵 🍉 🛇 🔲 🛜 🕪)) 99% [7]) pon. 19:52 🔍 ≔
			AntConc 3.4.3m (M	acintosh OS X) 2014
Corpus Files strategia_augustow.txt			Concordance Concordance Plot File	View Clusters/N-Grams Collocates Word List Keyword List
	Word Types:	1947 Word Tokens: 5515	Search Hits: 0	Lower West French
	Hank Fred	Lemma		Lemma word Form(s)
	1 12	9 augustow		augustowa 49 augustowem 1 augustowie 16 augustow 63
	2 61	miasto-panstwo		miast 5 miasta 54 miastami 2
	3 57	markowie		marki 57
	4 57	projekt		projekt 26 projektačni i projektu 19 projekty 2 projektow 9
	5 43	wczasowo-turystyczny		turystyczne 3 turystyczne 12 turystycznego 4 turystycznej 10 turystycznych 11 tury
	7 40	offective		produkcie s produkci iš produkcami i produkcu s produkcy s produkcow s
	8 40	launching		
	9 35	augustowski		augustowska 7 gugustowski 14 gugustowskich 4 gugustowskiej 6 gugustowskim 3 gugust
	10 30	nozycionowanie		nozvcjonowania 13 nozvcjonowania 15 nozvcjonowaniem 2
	11 30	promocyiny		promocying 1 promocyine 9 promocyinego 3 promocyinej 1 promocyiny 2 promocyinych 1
	12 29	komunikacia		komunikacja 10 komunikacji 11 komunikacja 1 komunikacje 7
	13 27	emo 🚽		emo 27
	14 26	działanie		działania 7 działaniami 1 działaniem 1 działań 17
	15 25	budować		budowania 2 budowanie 12 budować 5 budowała 1 buduje 4 budują 1
	16 25	należy		należałoby 1 należy 24
	17 25	rekomendować		rekomendowana 1 rekomendowane 5 rekomendowanego 1 rekomendowanych 1 rekomenduje 4
	18 24	impreza		imprez 5 impreza 9 imprezach 2 imprezami 1 imprezy 7
	19 24	miejsce		miejsc 2 miejsca 14 miejscach 1 miejsce 3 miejscem 3 miejscom 1
	20 24	np		np 24
	21 24	turystyka		turystyka 14 turystyki 6 turystyką 2 turystykę 2
	22 23	promocja		promocja 2 promocji 17 promocją 1 promocję 2 promocyj 1
	23 21	etap		etap 8 etapie 6 etapu 2 etapy 3 etapów 2
	24 21	kluczowy		kluczowe 12 kluczowych 5 kluczowym 4
	25 21	rekLamy		reklam 6 reklamy 15
	Search Term	Vords Case Regex	Hit Location	
	emo	Advance	d Search Only 0	
	Start	Stop Sort	Lemma List 🗸 Loaded	
Total No.	Sort by	Iver order		
1 Files Processed	Sort by Freq			Clone Results

Program przeniesie nas do zakładki **Concordance**, w której można podejrzeć kontekst wyrazu. Nasz wyraz jest fragmentem nazwy *EMO EFFECTIVE LAUNCHING*. Tak nazywa się firma projektująca strategię, która na każdej stronie dokumentu zaznaczyła swoją obecność. Więcej o tej opcji znajduje się w rozdziale 6.



Aby móc wyeksportować plik z frekwencjami wyrazów do arkusza kalkulacyjnego należy wrócić do zakładki **Word List**, wybrać z menu głównego **File** i **Save Output to Text File.** Następnie zapisujemy plik z wynikami.



Plik wynikowy jest plikiem tekstowym, który bardzo łatwo można otworzyć w arkuszu kalkulacyjnym. Do tego celu polecam użycie **programu Calc z pakietu Libreoffce**. W programie Calc otwieramy nasz plik i pojawia się opcja jego importu. Należy ustawić opcję kodowania znaków na **UTF-8**, oraz w opcjach separatora zaznaczyć - **Tabulator** i **Przecinek**. Następnie można już dać **OK**.

	AntConc 3.4.	am (iviacintosh OS X) 2014				
s Files	Concordance Concordance Plot	File View Clusters/N-Gran	ms Collocates \	Nord List Keyword List		
	Importuj tekst - [frekwencja_augustów.txt]					
· · · · ·	=/ Importui		BE A			
	Zestew znoków (Lisianda (LITE 0)			ug	ustów 63	
Liberation Sans	V 11	<u> </u>	.a. F			
	Język Domyślny - Polski	•				
¥				📼 🛌		turvstvcznych 11 tu
A	B Od wiersza 1		Н			duktów 9
	Oncie constatora				+	
				ść	Rodzaj	
	Stand-ozerokośc Torodzielony			B	Obrazek PNG	augustowskim 3 augu
	🥤 🔽 Tabulator 🔽 Przecinek 🔵 Średnik 🗌 Spacja	Inny		B	Obrazek PNG	
				B	Obrazek PNG	yjny 2 promocyjnych
	Scal separatory S	eparator tekstu 🛛 🚬			katalog	
				.B	Obrazek PNG	
	Inne opcje			.B	Obrazek PNG	
	Pole w cudzysłowie jako tekst Identyfikuj liczby s	specjalne		.B	Obrazek PNG	1
	Pola			.8	Obrazek PNG	and a selected in
	Fola			.D B	Obrazek PNG	nych 1 rekomenduje
	Typ kolumny			B	Obrazek PNG	. 1
	Standardowe StandardovStandardowe	Standardowe St		B	Obrazek PNG	1
	1 #Word Types: 1947			B	tekst	
	3 #Search Hits: 0					
	4 1 129 augustów	augustowa 49				
	6 3 57 markowie	marki 57				
	7 4 57 projekt	projekt 26				
	9 6 43 wczasowo-turystyczny	produkcie 3				
N Arkuez1						
AIRUSEL (**						
Znajdź	ОК	Anuluj Pomoc				
	Sort by Invert Order					
	Sort by Freq V					Clone Resu

5. Słowo w kontekście.

Concordance jest domyślną zakładka programu **Antconc**. Otwiera się ona natychmiast po uruchomieniu. Dostępne tu funkcje mogą służyć jako narzędzia do weryfikacji automatycznej analizy frekwencyjnej, mogą pomóc w poprawianiu błędów. Badacz posiada pełną kontrolę nad procesem. Natomiast najciekawszą funkcją jest zaawansowany **moduł wyszukiwania pełnotekstowego**. Jego funkcjonalność zostanie zademonstrowana na przykładzie słowa *kultura*.

Po wpisaniu słowa - *kultura* i uruchomieniu przycisku **Start**- w głównym okienku z wynikami pojawia się wyszukiwane słowo w kontekście tekstu.



Moduł wyszukiwania nie rozróżnia wielkich i małych liter. Dodatkowo w swym działaniu program pozwala na stosowanie znaków globalnych *. Tym razem wyszukamy słowo *kultura*, a także *kulturalny, kulturalne, kulturalnego*. Umieszczenie znaku globalnego po słowie *kultura** - zadziała w taki sposób, że program wszystkie słowa zaczynające się do tego rdzenia

🗯 AntConc	File Settings	Help 😵 🗋 🧼 🛇 🛜 🕪 🖇	2% 🔳 sob. 14:10 🔍 😑
		AntConc 3.4.3m (Macintosh OS X) 2014	
Corpus Files strategia_augustow.txt	Concords	Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List	
	Hit		File
	1	. I. SUBPRODUKT: TURYSTYKA 10 1.1.2. SUBPRODUKT: KULTURA 12 1.1.3. DZIAŁANTA PROMOCYJNE MIASTA AUGU	strategia_al
	2	. PROJEKT "TURYSTYKAŻYWIOŁOWA' 42 3.3. SUBPRODUKT KULTURALNY 43 3.3.1. PROJEKT "TRZYŻYWIOŁY KULTURY"	strategia_aı
	3	oferty m.in. turystycznej, kulturalnej i gospodarczej oraz przeglądzie	strategia_ai
	4	konne • Szlaki piesze Subprodukty kulturalno - sportowe • Mistrzostwa Polski w	strategia_au
	5	dla promowania marki. 1.1.2. Subprodukt: kultura Atrakcyjność Augustowa podnoszą również	strategia_aı
	6	doty- czących walorów turystyczno- kulturalnych miasta i regionu, a	strategia_aı
	7	sezonu. Cele dla subproduktu kulturalnego 1. Budowanie percepcji miejsca presti	strategia_aı
	8	powstające obiekty uzdrowiskowe. 3. Produkt kulturalny – należy stworzyć mocny produkt	strategia_aı
	9	 należy stworzyć mocny produkt kulturalny, który poza funkcją rozrywkową 	strategia_aı
	10	nowych turystów. Mocny produkt kulturalny powinien stać się kolejną	strategia_au
	11	- zydencji, która uszlachetnia. Produkt kulturalny traktujemy w sposób ela-	strategia_au
	12	oferty turystycznej, inwestycyjnej i kulturalnej. Kluczowe założenie tego pomysłu	strategia_au
	13	trzy kluczowe subprodukty: turystyczny, kulturalny i gospodarczy. Zostaną one	strategia_ai
	14	najblizszych kilku lat. 3.3. Subprodukt kulturalny Naszym zdaniem produkt kulturalny	strategia_ai
	15	kulturalny Naszym zdaniem produkt kulturalny na wysokim poziomie, będący	strategia_al
	10	. Rekomendujemy zatem stworzenie produktu kulturalnego, który stanie się flagowym	strategia_al
	17	stanie się riagowym wyaarzeniem kulturalnym na mapie polski a	strategia_a
	10	". W KONTEK- SCLE Projektu Kulturalnego derintujemy "zywior" jako zywiorowość	strategia_a
	20	Cole di submodulta, kulturatine jaugustova, poprzez ktorą interzatoby	strategia a
	21	rzeczowy konkursy) Dia subproduktu kulturatiego i badomante perceptji kagustowa jako	strategia a
	22	(eventualnie) - bilboardy kaj Subarodukt - kulturalny Tay Żywiały Kultury Posso -	strategia a
	Search Te	arm Words Case Regex Search Window Size	
	kultura*	Advanced 50 C	
Total No. 1	Kwic Sor	t 11 TR C V Level 2 2R C V Level 3 3R C	Clone Results
Files Processed	-		

Dodatkowo program pozwala na sortowanie wyników alfabetycznie. Domyślnie ustawione jest sortowanie słów po prawej stronie. Na pierwszym poziomie sortowane jest pierwsze słowo po prawej, następnie drugie i wreszcie trzecie. Kolejność tą można zmienić, jak również kierunek sortowania, w taki sposób, że będą sortowane słowa po lewej stronie.

AntConc File	Settings Help		🛜 🜓)) 83% 🔳 sob. 14:27 🔍
•		AntConc 3.4.3m (Macintosh OS X) 2014	
orpus Files		Concertance Concertance Dist Ella View Clustere/N-Crome Collectee Word List Keyword List	
trategia_augustow.txt		Concordance Concordance Piot Pine View Clusters/N-Grams Collocates Word List Reyword List	
	Concordance Hits 29 Hit KWIC		File
	1	dla papaganania marki 112 Subaradukt, kultura Ataglandaté Augustana padagang démiat	stratagia a
	2	ata promowania marki. 1.1.2. Subprodukt: kultura Atrakcyjnosć Augustowa podnoszą rownież	strategia a
	3	turystow - wizy- towkę orerty kulturalneg Augustowa, poprzez którą należatoby	strategia a
	4	cete tru subprouktu kulturunego i. Budowante percepcji Augustowa jako	strategia a
	5	" W kontaka faja projektu kulturunego i ofinijuma periopoji miejsku prest	strategia a
	6	T SUBPRODUKT TUPYSTYKA 10 1 1 2 SUBPRODUKT KULTUR 12 1 3 DZTA ANTA PROMOCYINE MTASTA AUG	strategia a
	7	oferty min invisionali, kulturalnej i ospodarczej oraz przedadzie	strategia a
	8	tray kluczowe submodukty turystyczny kulturalny i gospodarczy Jostana ope	strategia a
	9	oferty turystycznej i westycznej i kulturalnej Kluczowe założenie tego powysłu	strategia a
	10	- noleży stworzyć mocry produkt kulturalny, który poza funkcji opznywawa	strategia al
	11	Rekomendijemy zitem stvorzenje produktu, kulturalnego, który stanje sie flagovym	strategia al
	12	aktywnym wypoczynkiem, odkrywaniem zabytków kultury lub pomików przyrody. • Kanał	strategia al
	13	doty - czących walorów turystyczno- kulturalnych miasta i regionu, a	strategia al
	14	stanie sie flagowym wydarzeniem kulturalnym na manie Polski a	strategia_au
	15	kulturalny Naszym zdaniem produkt kulturalny na wysokim poziomie, bedacy	strategia_au
	16	powstające obiekty uzdrowiskowe. 3. Produkt kulturalny - należy stworzyć mocny produkt	strategia_ai
	17	najbliższych kilku lat. 3.3. Subprodukt kulturalny Naszym zdaniem produkt kulturalny	strategia_au
	18	jak Trzy Ży- wioty Kultury, Oprócz reklam w radiowej	strategia_au
	19	subproduktu kulturalnego: • Trzy Żywioły Kultury. Podobnie jak w poprzednich	strategia_aı
	20	nowych turystów. Mocny produkt kulturalny powinien stać się kolejną	strategia_au
	21	3 Subprodukt kulturalny Trzy Żywioty Kultury Prasa - Przekrój, Newsweek; Radio -	strategia_au
	22	festiwa- lu Trzy Żywioły Kultury. Projekt powiniem opierać sie	strategia_au
	23	PRODUKT KULTURALNY 43 3.3.1. PROJEKT "TRZYŻYWIOŁY KULTURY" 44 3.3.2. PROJEKT "REZYDENCJA KRÓLEWSKA"	strategia_au
	24	. PROJEKT "TURYSTYKAŻYWIOŁOWA' 42 3.3. SUBPRODUKT KULTURALNY 43 3.3.1. PROJEKT "TRZYŻYWIOŁY KULTURY"	strategia_au
	25	konne • Szlaki piesze Subprodukty kulturalno - sportowe • Mistrzostwa Polski w	strategia_au
	26	- zydencji, która uszlachetnia. Produkt kulturalny traktujemy w sposób ela-	strategia_aı
	Search Term 🔽 Words	Case Regex Search Window Size	
	kultur*	Advanced 50	
	Otart Otar		
	Start Stop	Sort	
al No.		aval 2 28 2 1 aval 3 28	Clone Beaulte
			Cione Results

Zmiana trybu sortowania z 1R (1 prawy) - 2R (2 prawy)- 3R (3 prawy) na 1L (1 lewy) - 2L (2 lewy) - 3L (3 lewy) ujawnia ciekawszą cechę tekstu. Na 29 wystąpień słowa *kultura* - 15 razy pojawia się ono w kontekście *produktu* i *subproduktu*.

