Zastosowanie kolokacji językowych w badaniach ilościowych

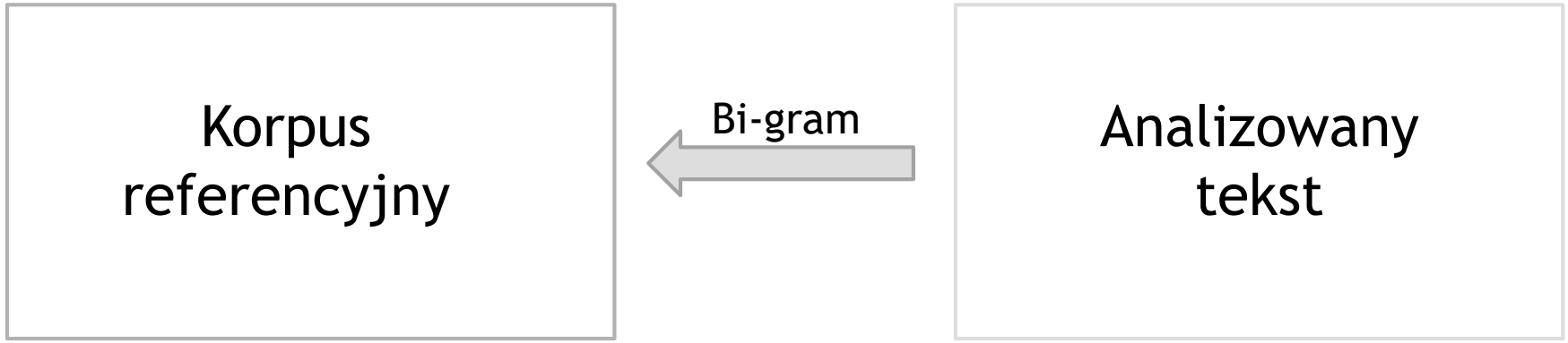
mgr inż. Agnieszka Wołk

Instytut Badań Literackich PAN

Polsko-Japońska Akademia Technik Komputerowych

Inspiracja

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.



Profil Collgram

Profil Collgram

MI – Mutual Information
Punktowa wzajemna informacja

$$MI(x,y) = \log_2 \left(\frac{N \cdot f(x,y)}{f(x) \cdot f(y)} \right)$$

T-score

$$T(x,y) = \frac{f(x,y) - \frac{f(x) \cdot f(y)}{N}}{\sqrt{f(x,y)}}$$

Liczba jednostek idiosynkratycznych

Bazuje na częstotliwości bigramów

LONGDALE

Longitudinal Database of Learner English

Badania wzdłużne

Styczeń 2008

Język ucznia – angielski

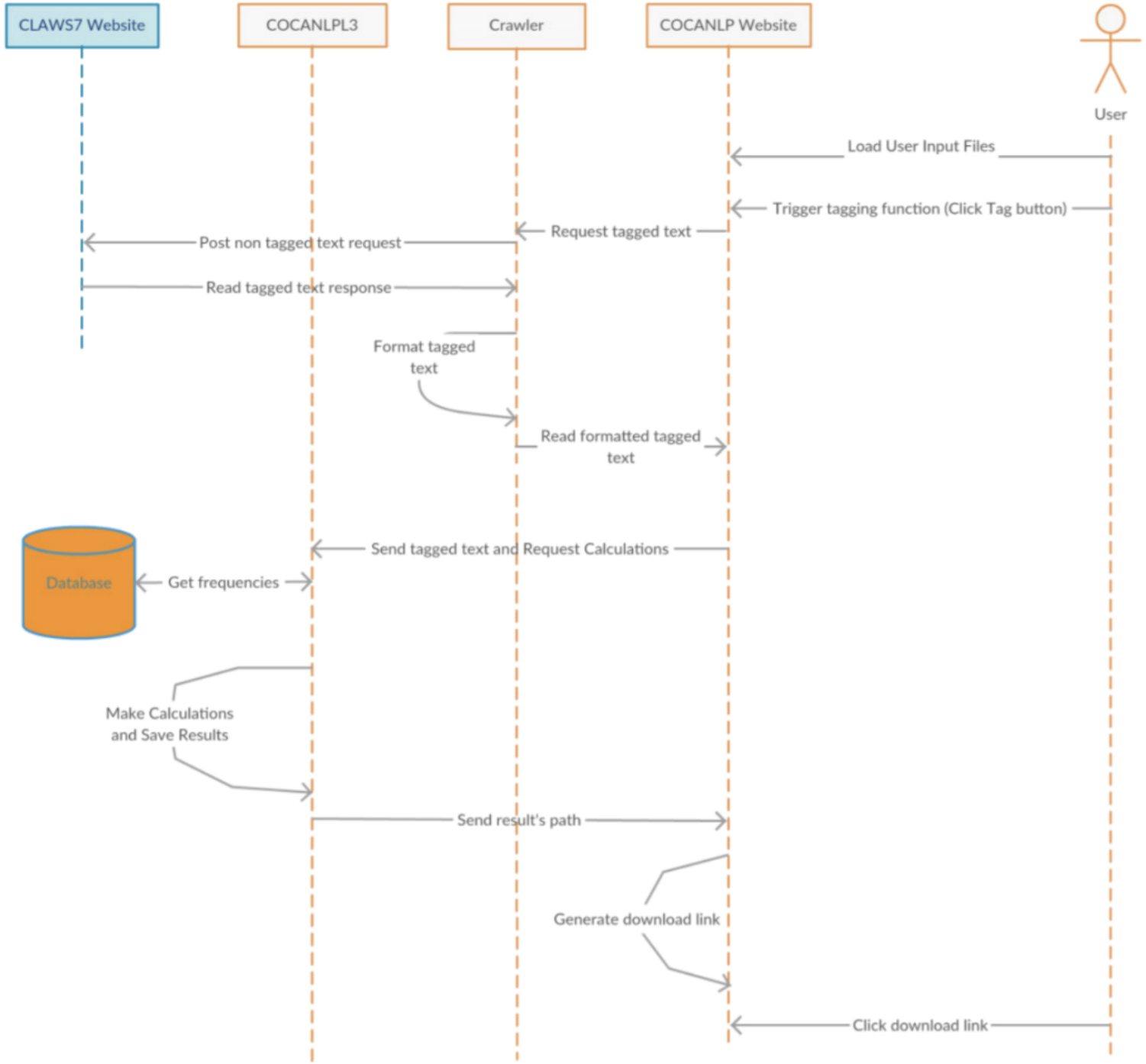
Postępy ucznia są odnotowywane na przestrzeni minimum 3 lat

Baza danych.

Université catholique de Louvain

Aplikacja pierwotny schemat

- Pionowa linia przerywana – czas aktywności i działania elementu
- Strzałka – akcja wywołana przez jeden obiekt wobec drugiego lub samego siebie
- Niebieski element – obiekt poza rozwiązaniem



Aplikacja

- .NET - C#
- WWW - Microsoft IIS.
- warstwa przetwarzania języka naturalnego - Python - NLTK (Natural Language Toolkit).
- MongoDB.
- wykluczenie z analizy słów nie znajdujących się w COCA - błędy ortograficzne, nieistniejące słowa, obce słowa i mniej pospolite nazwy własne.

NLTK

Biblioteki i programy do NLP

Steven Bird, Edward Loper

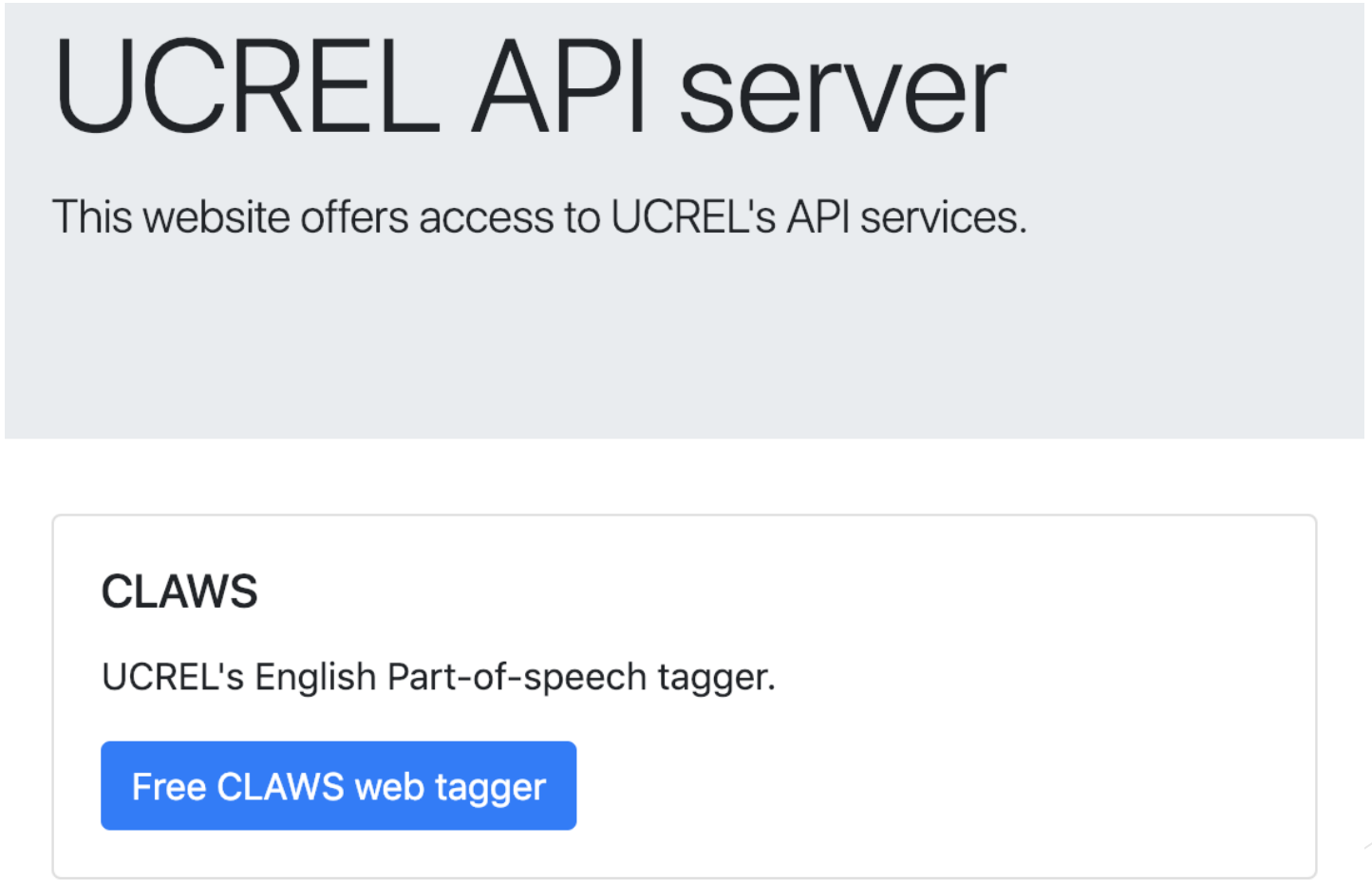
University of Pennsylvania

2001

Obliczenia MI i t-score

CLAWS

- Tagger POS
- He’s = he is
  or he has



PREPRO

- Skrypty
  - Tokenizacja
  - Truecasing
  - Normalizacja
- http://data.statmt.org/ngrams/prepro.tgz

COCA

- Korpus współczesnego amerykańskiego języka angielskiego
- AKTUALNIE już 560 mln słów
- Mark Davies profesor lingwistyki korpusowej na uniwersytecie Brigham Young.

OPENSUBTITLES

- Opensource
- Napisy filmowe
- PL-EN
  - 27 mln zdań
  - 323mln słów
- EN-PL
  - 23mln zdań
  - 303mln słów

Aplikacja – dane wejściowe

- Pliki txt wstępnie przygotowane i oczyszczone
- Osobno
- Opcja ” Exclude non-COCA words ” pozwala na wyłączenie z analizy słów, których nie ma w korpusie referencyjnym COCA

Zautomatyzowane procesy

1) Tokenizacja: każdy tekst ucznia jest tokenizowany, a następnie oznaczany znacznikami POS w celu identyfikacji odpowiednich nazw i znaków interpunkcyjnych.

2) ekstrakcja n-gramów: wszystkie n-gramy są ekstrahowane z każdego tekstu L2. Znaki interpunkcyjne i dowolna sekwencja znaków, która nie odpowiada słowu, przerywa n-gramy.

3) n-gramy które zawierają słowo zidentyfikowane przez tagger jako nazwę własną lub numer są wyłączone z analiz

4)Obliczanie wyników powiązania: każdy bi-gram jest wyszukiwany w korpusie referencyjnym i przypisywany do MI i t-score obliczanych za pomocą wzorów przedstawionych w pracy.

5) Obliczanie profili CollGram: 3 wskaźniki – średnia MI, średni wynik t, a także proporcja nieobecnych bi-gramów

Usprawnienia

3-gram 4-gram	Wsparcie dla innych języków	word2vec
Większe modele językowe	Miara Dice $Dice(x,y) = \log_2 \left(\frac{2 \cdot f(x,y)}{f(x) + f(y)} \right)$	Lexical Gravity $G(x,y) = \log \left(\frac{f(x,y)}{g(x)} \cdot \frac{f(x,y)}{g'(y)} \right)$

Współpraca z
Yves Bestgen
Universite catholique de Louvain

Daudaravičius, V., & Marcinkevičienė, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, *9*(2), 321-348.

Miara DICE

- Współwystępowanie słów lub grupy słów

$$Dice(x,y) = \log_2 \left(\frac{2 \cdot f(x,y)}{f(x) + f(y)} \right)$$

  - f(x,y) to częstotliwość współwystępowania x i y,
  - f(x) i f(y) to częstotliwości występowania x i y w dowolnym miejscu w tekście.

- Jeśli x i y mają tendencję do występowania w połączniu, ich wynik DICE będzie wysoki.

Lexical Gravity

- Ocena możliwości połączenia dwóch wyrazów w tekście.
- Jeśli pierwsze słowo x jest używane częściej niż oczekiwano przed drugim słowem y, a drugie słowo y jest używane częściej niż oczekiwano po pierwszym słowie x, wówczas x i y tworzą kolokację.

$$G(x,y) = \log\left(\frac{f(x,y)}{g(x)} \cdot \frac{f(x,y)}{g'(y)}\right)$$

- f(x, y) to częstotliwość pary słów x i y w korpusie; g(x) to współczynnik różnorodności słów po prawej stronie x; g’(y) to współczynnik różnorodności słów po lewej stronie y.

Word2vec

- Wektorowa reprezentacja słów
- Przewiduje wzajemne sąsiedztwo słów, nie jest ważny szyk zdania
- Wynik działania treningu word2vec to wektory słów oddające podobny kontekst występowania słów
- Wychwytuje sensowne prawidłowości składniowe i semantyczne, regularności - odbierane jako stałe przesunięcia pomiędzy parami słów.
- Reprezentacja wektorowa – b.dobra w odpowiadaniu na pytania w oparciu o analogie. (np. a do b jak c do ?).

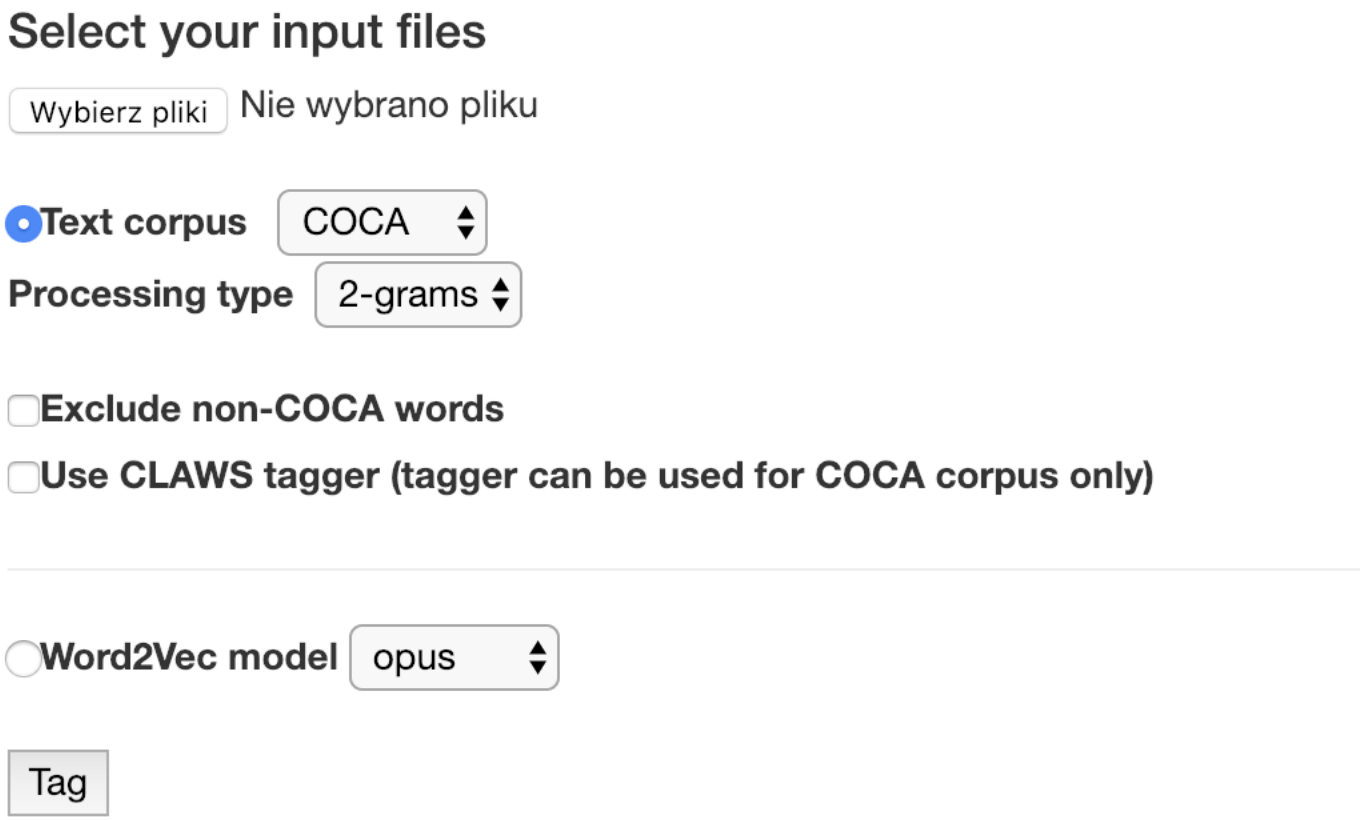
Na przykład: *mężczyzna ma się do kobiety jak wujek do ? (cioci)*.

Tego typu analogie określa się za pomocą metody przesunięcia wektorowego opartej o odległość kosinusową.

Zastosowanie word2vec w NLP

- Ulepszanie tłumaczenia maszynowego
- Wyszukiwanie informacji
- Wytrenowanie modelu word2vec na korpusie referencyjnym i wyciągnięcie z niego częstotliwości bi-gramów
- odejście od reprezentacji ilościowej na rzecz wektorowej i analiza odległości kosinusowej w tekstach ucznia w stosunku do tekstu referencyjnego

DEMO APLIKACJI



http://collgram.pja.edu.pl/

Select your input files

Wybierz pliki Nie wybrano pliku

Text corpus ✓ COCA
opus_en
opus_pl

Processing ty

Select your input files

Wybierz pliki Nie wybrano pliku

Text corpus COCA

Processing type 2-grams

Exclude non-COCA words

Use CLAWS tagger (tagger can be used for COCA corpus only)

Word2Vec mode ✓ opus
opus_en

Tag

COCA Parser Home Tasks About Contact

Tasks

Task Id	Created at	Status	File count	Results
07f18429-cdad-401e-95fa-46ed5e3d8d4e	2019-04-05 14:56:57	Finished	1	Download
00270fb6-c698-41f4-a909-ec59e02731b5	2019-04-03 10:57:59	Finished	1	Download
85fa4720-a25e-4ff9-8548-4508c29af2c7	2019-04-02 20:00:45	Finished	1	Download
b606a805-0e9b-4da2-9958-519e33d44ad3	2019-04-02 10:47:09	Finished	1	Download
9b1bc7a0-b192-41f9-854d-fdccdad26ab0	2019-04-02 09:25:51	Finished	1	Download
d99c3fd0-d36f-4685-b1b3-eec4f8cea6b0	2019-03-29 07:28:48	Created	1	Download
c55457f9-f5d9-484c-8694-6b0591525991	2019-03-29 07:27:40	Created	1	Download
f0b383bd-b3da-44a6-a0b4-4328fcfe5afd	2019-03-29 07:27:19	Created	1	Download
a8a7c34c-7d0a-4966-809d-37ccfb109794	2019-03-23 03:16:14	Created	1	Download

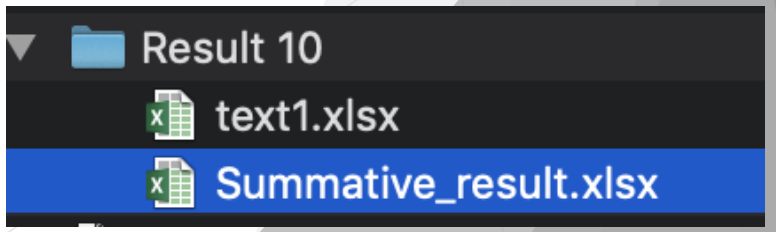
Aplikacja – dane wyjściowe

.ZIP

.XLSX

inputfile.xlsx

Summative_result.xlsx



	A	B	C	D	E	F	G	H	I	J	K
1	bigram typ	freq in text	freq in COCA	mean freq in COCA	MI	MI>3	t	t>2	dice	gravity	
2	the left	6	16851	39,6494117647059	0,6388575		46,444302	*	-6,568028	-4,20944797394969	
3	left of	2	3541	8,33176470588235	-0,541812		-27,10637		-7,395220	-5,24236492694466	
4	who refuse	2	337	0,792941176470588	4,2058771	*	17,377037	*	-7,410942	-4,83466385131303	
5	refuse thos	1	1	0,0023529411764705	-2,758436		-5		-12,24885	-9,45876107661153	
6	those living	1	485	1,14117647058824	2,4811845		18,117656	*	-6,245894	-5,1957362054192	
7	living with	2	4168	9,80705882352941	2,7575802		55,018545	*	-5,886612	-4,17834685719167	
8	with hiv	2	1010	2,37647058823529	4,0538064	*	29,892546	*	-7,276729	-4,40375021559599	
9	hiv are	2	35	0,0823529411764706	-0,647872		-3,211586		-10,53601	-7,27934999286913	
10	are a	5	58915	138,623529411765	-0,030718		-5,219917		-4,653382	-3,86741012802644	
11	a perfect	1	7654	18,0094117647059	3,1409536	*	77,577114	*	-6,459354	-3,79897921874828	
12	perfect exa	1	618	1,45411764705882	5,8804643	*	24,457346	*	-4,835446	-4,06720429234706	
13	you talk	1	6052	14,24	2,4412344		63,474841	*	-5,664951	-4,11160520717331	
14	talk about	6	48718	114,630588235294	6,9570275	*	218,94555	*	-2,606707	-1,84642550251521	
15	about whe	1	1976	4,64941176470588	-0,530223		-19,72904		-6,325491	-5,49221502623952	
16	when nigel	1	2	0,0047058823529411	-0,160744		0		-12,45805	-8,37574786440364	
17	said that	8	45662	107,44	1,7347516		149,48521	*	-4,288155	-3,44662171540089	
18	that the	74	399093	939,042352941176	0,3445283		134,20435	*	-3,606876	-2,92360145933155	
19	the nhs	2	220	0,517647058823529	3,0030669	*	13,012057	*	-10,89850	-5,38195302071092	
20	nhs should	1	1	0,0023529411764705	1,3692948		1		-12,03238	-8,21619011121853	
21	should not	6	17602	41,4164705882353	3,6422176	*	122,05239	*	-4,098167	-3,28641015645716	
22	not be	11	73212	172,263529411765	3,1181207	*	239,41414	*	-3,254323	-2,82515878992594	
23	be used	3	22258	52,3717647058824	4,4701565	*	142,46153	*	-3,909697	-2,9352515735165	
24	used to	2	81621	192,049411764706	3,8808173	*	266,30250	*	-4,236591	-2,54834458944225	
25	to pay	8	34595	81,4	3,9356883	*	173,84659	*	-5,083763	-2,9046153589535	
26	pay for	4	16702	39,2988235294118	4,4620850	*	123,37872	*	-4,734180	-3,06239900635827	
27	for those	2	20441	48,0964705882353	2,5107081		117,89019	*	-4,613921	-3,56208833927228	
28	those who	25	57032	134,192941176471	5,7361827	*	234,33744	*	-2,588559	-2,14550726029402	
29	who come	2	2432	5,72235294117647	1,6621492		33,742055	*	-5,664214	-4,74206858964432	
30	come to	2	47273	111,230588235294	2,6208423		182,07779	*	-4,789998	-3,16482361861592	
31	to the	76	1055357	2483,19294117647	0,7662966		423,32773	*	-2,804903	-2,37431124802532	
32	the uk	4	1078	2,53647058823529	3,1887757	*	29,238955	*	-9,309333	-4,63585301393303	
33	uk to	2	15	0,0352941176470588	-1,875280		-10,32795		-12,81957	-8,0168134510694	

	A	B	C	D	E	F	G	H	I	J	K
1	file name	word tokens	bigram tokens	bigram types	mean freq tokens	mean freq types	mean MI tokens	mean MI types	mean t tokens	mean t types	
2	newsdiscus	24220	20036	12706	252,2620257653	19,71674188201	2,16277451995(	1,9046960769	111,0236721878	21,8224271925454	
3											
4											
5											
6											
7											

Zastosowania

Ocena kompetencji językowych

FLA i SLA – krótkie porównanie

Tabela 1. Krótkie porównanie FLA i SLA

FLA	SLA
Instynkt, wyzwalany przez narodziny	Osobisty wybór, wymagana motywacja
Bardzo szybki	Zmienny, jednak nie tak szybki jak FLA
Kompletny	Nigdy nie będzie tak dobry jak natywny, jednak można osiągnąć bardzo dobre wyniki
Naturalny	Naturalny lub prowadzony (dla języków syntetycznych wymagana jest znajomość gramatyki)

Unsupervised tool for quantification of progress in L2 English phraseological

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 383–388
DOI: 10.15439/2017F433
ISSN 2300-5963 ACSIS, Vol. 11

Unsupervised tool for quantification of progress in L2 English phraseological

Krzysztof Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kwolk@pja.edu.pl

Agnieszka Wołk
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: awolk@pja.edu.pl

Krzysztof Marasek
Polish-Japanese Academy of Information Technology,
ul. Koszykowa 86, 02-008
Warszawa, Poland
Email: kmarasek@pja.edu.pl

Abstract **- This study aimed to aid the enormous effort required to analyze phraseological writing competence by developing an automatic evaluation tool for texts. We attempted to measure both second language (L2) writing proficiency and text quality. In our research, we adapted the CollGram technique that searches a reference corpus to determine the frequency of each pair of tokens (bi-grams) and calculates the t-score and related information. We used the Level 3 Corpus of Contemporary American English as a reference corpus. Our solution performed well in writing evaluation and is freely available as a web service or as source for other researchers.**

I. Introduction

A person's second language, or L2, is a language that is not the native language of the speaker but is used in the locale of that person. In contrast, a foreign language is a language that is

It is important to analyze the role of corpus linguistic studies in the grading of L2 writing. In such grading, it is essential to analyze the writing based on functional skills and the independent construction of written text to communicate in a purposeful context. A human writer cannot be used to demonstrate the requirements of the standards, as this does not meet the requirement for independence. In writing assessment, we should consider whether or not information and ideas were presented concisely, logically, and persuasively. It is also important to determine whether or not a writer clearly presented information on complex subjects, used a range of writing styles for different purposes, and employed a range of sentence structures, including complex sentences and paragraphs, to effectively organize their written communication. We should also evaluate the accuracy of punctuation in written text using commas, apostrophes, and quotation marks. Lastly, written

FedCSIS
Praga, Czechy
3 - 6 Września 2017

Miara TER

- Miara TER została zaprojektowana z myślą o tłumaczeniu maszynowym. Głównymi jej założeniami było zapewnienie intuicyjności przy jednoczesnym wymogu mniejszej próbki testowej niż przy innych technikach oceny tłumaczenia maszynowego.
- Kalkuluje ona ilość edycji, które należy zrobić, aby tekst przetłumaczony stał się taki jak referencyjny w kontekście płynności i semantyki wypowiedzi.

$$TER = \frac{E}{W_R}$$

Grupa badawcza - PL

- 50 obcokrajowców
- 16-39 lat
- 67% to kobiety
- od 1,000 do 1,200 słów
- „Moje wspomnienia z najlepszej wycieczki w moim życiu.”
- Sprawdzone przez polonistę – zaznaczenie błędów
- Wyliczone TER

Współczynnik korelacji pomiędzy miarą TER i MI: 0.88744429 (dość silna zależność)

Analiza Collgram w wykrywaniu wczesnych objawów depresji.

DLACZEGO COLLGRAM?

- Większość algorytmów skupia się głównie na tym, o czym jest tekst, przez co łatwo je oszukać
- Nie skupiamy się na tym o czym jest tekst, tylko jak jest napisany.

Badanie

- Fora depresja
- Artykuły o depresji
- Teksty pisane przez psychologów i psychiatrów

Najczęściej używane słowa

- absolutely
- all
- always
- complete
- completely
- constant
- constantly
- definitely
- entire
- ever
- every
- everyone
- everything
- full
- must
- never
- nothing
- totally
- whole

WNIOSKI

- WYŻSZY MI
- WYŻSZY T-SCORE
- MNIEJSZY ZASÓB SŁÓW
- Osoby z depresją używają też znacznie częściej zaimków w pierwszej osobie liczby pojedynczej taki jak „ja”, „moje”, „mnie”, „mi”, co sugeruje skupienie na samym sobie i brak zainteresowania otaczającym ich światem.

Potencjalne implikacje i zastosowania

Projekty PROMIS i HPO
(Human Phenotype Onthology)

- WUM
- Dodatkowy czynnik decyzyjny

WARSZAWSKI UNIWERSYTET MEDYCZNY

Polish Telemedicine Society
Polskie Towarzystwo Telemedycyny i e-Zdrowia

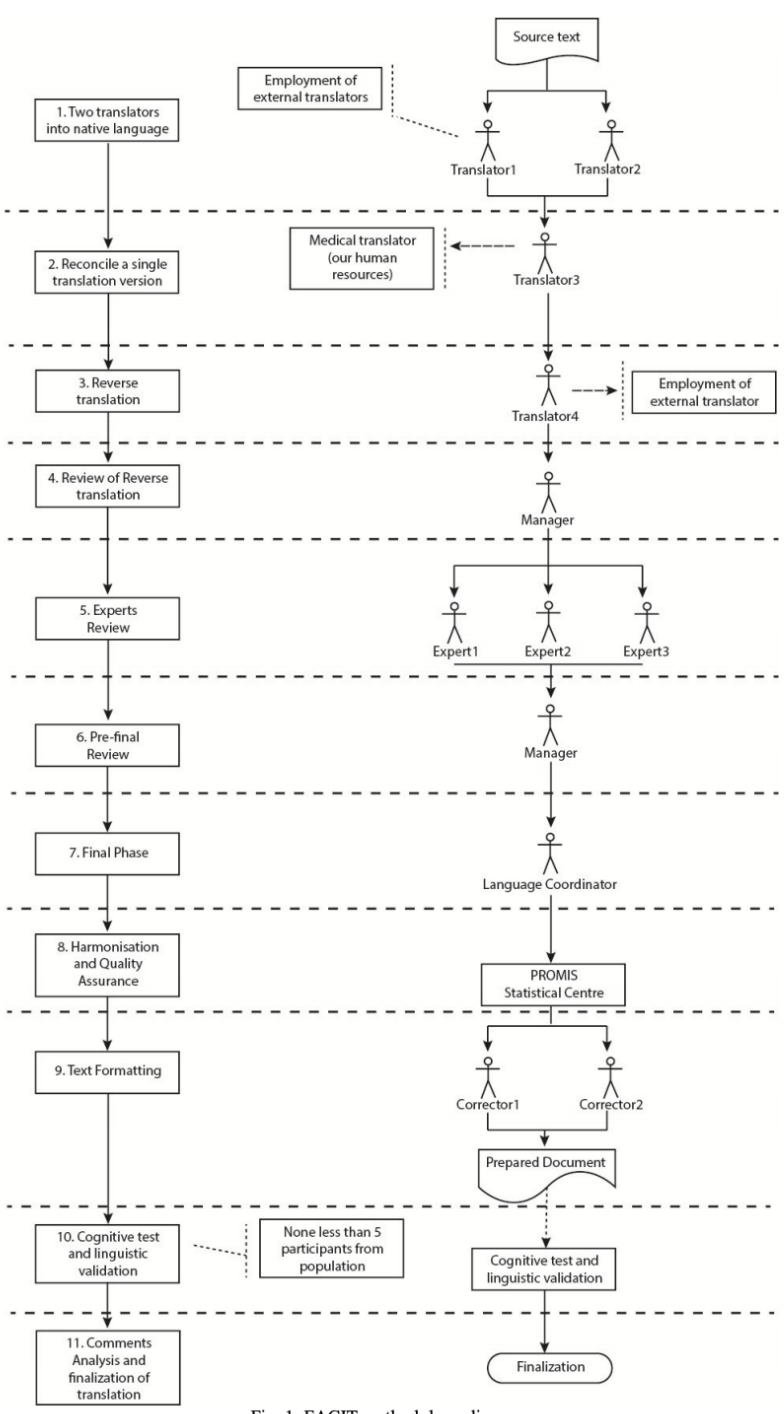


Fig. 1. FACIT methodology diagram.

RESPEAKING

- Miara NER

$$NERvalue = \frac{N - E - R}{N} * 100.$$

  - N - całkowita liczba słów w podpisie
  - E – błędy edycyjne
  - R – błędy w rozpoznawaniu ASR

WOŁK, Krzysztof; KORŽINEK, Danijel. Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. *arXiv preprint arXiv:1601.02789*, 2016.

OCENA TŁUMACZENIA MASZYNOWEGO

- Konkurs WMT (Conference on Machine Translation)

Stylometria

KORPUS UCZNIA
Czym jest korpus ucznia?

- Zawiera teksty pisane przez obcokrajowców uczących się danego języka
- Zawiera oceny i poprawy nativespeakera
- Nie ma dla języka polskiego.

Korpus ucznia język POLSKI

- Italki
- Lang8

Narzędzie dostępne pod adresem:

http://collgram.pja.edu.pl

Dziękuję za uwagę.

awolk@pja.edu.pl