



POLISH INTERPRETING CORPUS

Polish Interpreting Corpus:

jak korpus może pomóc w badaniach
nad procesem tłumaczenia
symultanicznego

Agnieszka Chmiel



UNIWERSYTET
IM. ADAMA MICKIEWICZA
W POZNANIU

Danijel Koržinek



POLSKO-JAPOŃSKA AKADEMIA
TECHNIK KOMPUTEROWYCH



N A R O D O W E C E N T R U M N A U K I

2018/30/E/HS2/00035 SONATA BIS

Ekstremalna regulacja języka: aktywacja i inhibicja
jako mechanizmy kontroli dwujęzycznej w tłumaczeniu konferencyjnym

Grupa badawcza



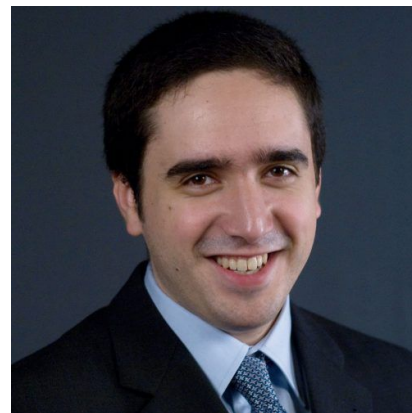
Agnieszka Chmiel
Wydział
Anglistyki
UAM



Marta
Kajzer-Wietrzny
Wydział Anglistyki
UAM



Przemysław
Janikowski
Instytut
Językoznawstwa UŚ



Danijel Koržinek
PJATK



Dariusz Jakubowski
Instytut
Językoznawstwa UŚ



Tłumaczenie symultaniczne

Pytania badawcze

1. Jak mechanizmy regulacji języków (aktywacja i inhibicja) funkcjonują w tłumaczeniu konferencyjnym (które wykonuje się w ekstremalnych warunkach: przy obciążeniu kognitywnym i obostrzeniach czasowych)?
2. Czy doświadczenie i szkolenie tłumaczeniowe poprawiają regulację języków w umyśle dwujęzycznym?

Metody badawcze

Korpus - dane autentyczne

Badania empiryczne
(quasi-eksperymentalne):

- tłumacze konferencyjni
- studenci specjalizacji
tłumaczeniowych na początku i
pod koniec szkolenia
- zadania nietłumaczeniowe,
tłumaczenie symultaniczne,
symultaniczne z tekstem, a vista
(okulograf)



Korpusy tłumaczenia ustnego

Rodzina EPIC

EPIC (Monti et al. 2005)

TIC (Kajzer-Wietrzny 2012)

EPTIC (Ferraresi i Bernardini 2019)

EPIC - Ghent (Defrancq et al. 2015)

Korpusy polsko-angielskie

Dumara (2015)

Bartłomiejczyk (2016)

podkorpus EPTIC (Kajzer-Wietrzny 2020)



PINC

Parlament Europejski

przemówienia eurodeputowanych i tłumaczenia (100-500 słów)

przemówienia czytane i mówione

styczeń 2009 - wrzesień 2010

dane eksperymentalne o tłumaczach (Chmiel 2016, 2018)

39 tłumaczy (EN-PL, PL-EN)



Oryginały polskie
232 przemówienia
40,168 tokenów

Tłumaczenia na angielski
232 przemówienia
41,283 tokeny



Oryginały angielskie
230 przemówień
54,090 tokenów

Tłumaczenia na polski
230 przemówień
41,283 tokeny

Tłumaczenia
na angielski
ok. 10,000 tokenów

PINC

POLISH INTERPRETING CORPUS

Przetwarzanie danych



Korpusowa analiza danych

Utworzenie korpusu mowy - kroki:

1. pozyskiwanie nagrań
 2. transkrypcja
 3. dopasowanie transkrypcji i nagrania, segmentacja
 4. rozpoznawanie głosów tłumaczy
 5. anotacja lingwistyczna
 6. dopasowanie dwujęzyczne
-



Pozyskiwanie nagrań

Pliki wideo

<https://www.europarl.europa.eu/>

wielokanałowe audio (dla różnych języków)

format MP4, audio AAC 44.1 kHz

konwersja do nieskompresowanego 16 kHz

Transkrypcje

wygładzanie wypowiedzi oryginału

tłumaczenia nie są takie same jak mowa



Wyrównanie tekstu i audio

Speech alignment

segmentacja na poziomie wyrazów

forced alignment i wierność transkrypcji

endpointing

EMU-webApp - weryfikacja i poprawianie!

Gitlab

backend danych

praca zespołowa

historia zmian





Identyfikacja głosów tłumaczy

X-vectors

embedding segmentów nagrań w przestrzeni głosów

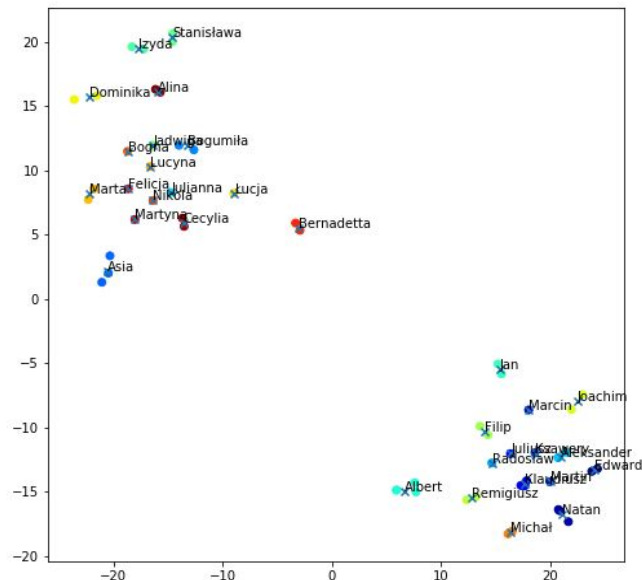
zbiory danych: development, enrollment/train, test

1 wektor na mówcę

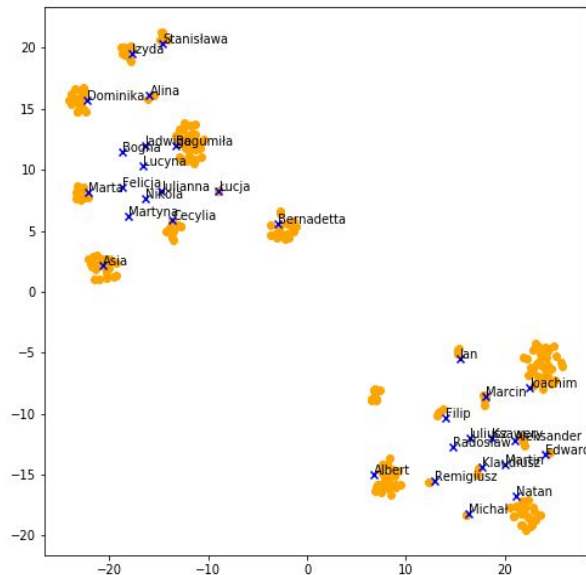
PLDA do określenia dopasowania mówca-plik

macierz odległości

Identyfikacja - trenowanie



Identyfikacja - rozpoznawanie





Anotacja lingwistyczna

Cechy lingwistyczne

tokenizacja, POS, dependency parsing
spaCy, NLTK, morefusz+concraft, KRNNT

Panie przewodniczący, chciałem zapytać pana o odbyty niespełna dwa tygodnie temu szczyt Unii...

częstości rzeczowników (korpus Europarl)



Dopasowanie dwujęzyczne

Alignment

hunalign

<http://www.intertext.cz/>

symgiza++

work in progress



Analiza korpusowa

EMU i R

wyszukiwanie segmentów - EQL

liczenie statystyk

ekstrakcja cech akustycznych

Badania korpusowe

1. Parametry płynności
2. Obciążenie kognitywne
3. Tłumaczenie kognatów
i homonimów

PINC

POLISH INTERPRETING CORPUS

Parametry płynności



Parametry płynności

Jaki jest poziom kompresji tekstu w tłumaczeniu?

Jak prędkość oryginału wpływa na tłumaczenie?

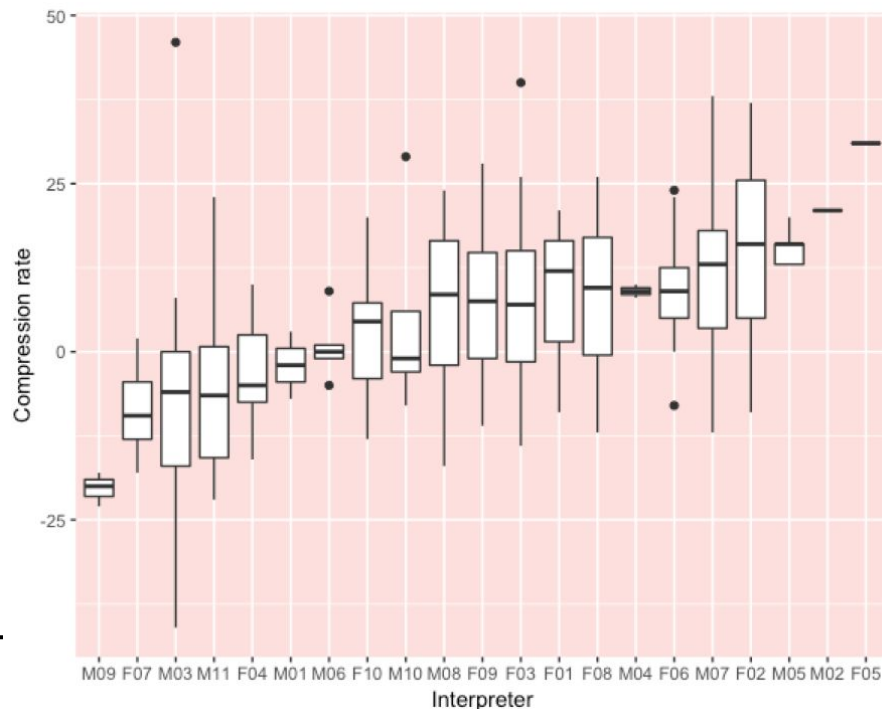
Czy oryginały i tłumaczenia różnią się pod względem płynności?

Kompresja tekstu w tłumaczeniu

$(\text{liczba tokenów w ST} - \text{liczba tokenów w TT}) * 100 / \text{liczba tokenów w ST}$
(Russo 2018)

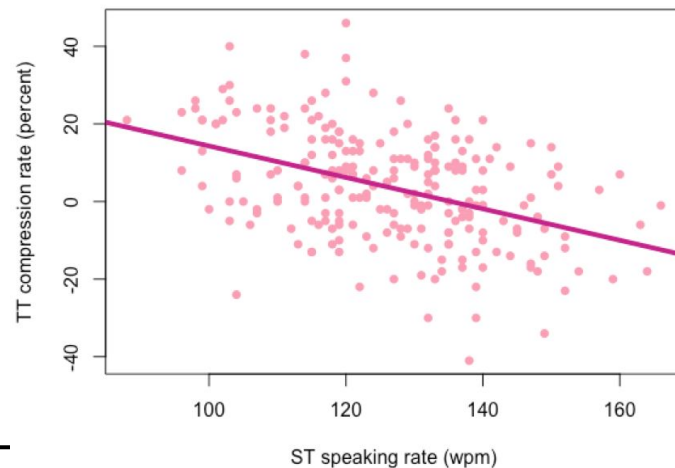
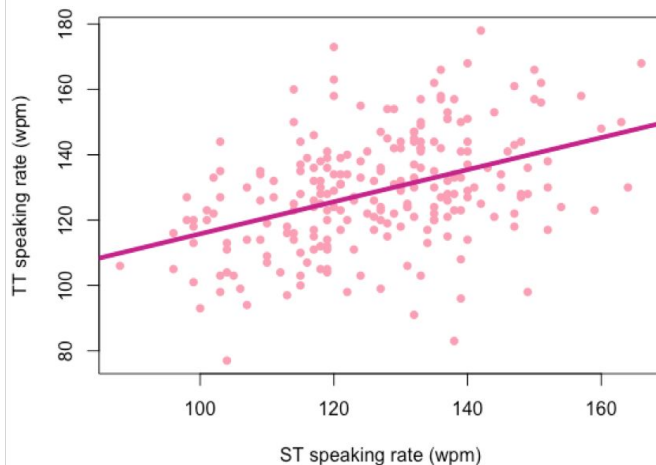
0: brak kompresji, ST=TT
<0: kompresja TT
>0: wydłużenie TT względem ST

PL-EN +3,6%



Prędkość oryginału a tłumaczenie

Im szybszy oryginał, tym szybszy tłumacz i większa kompresja



Różnice w płynności między ST i TT

Liczba i długość cichych i wypełnionych pauz

pauzy ciche >250ms, pauzy wypełnione - brak odcięcia

Długość nieprzerwanej wypowiedzi

wypowiedź bez pauz (w sylabach)

Proporcja wypowiedzi

ile procent długości tekstu stanowi artykulacja (a nie pauzy)



Parametr	ST	TT	t	p
Długość	1 min 22 s	1 min 24 s	-.28	=.78
Liczba cichych pauz/min	12.06	10.36	5.51	p<.001*
Średnia długość cichej pauzy (w ms)	487	626	-6.86	p<.001*
Liczba pauz wypełnionych/min	2.84	7.37	-11.10	p<.001*
Średnia długość pauzy wypełnionej (w ms)	613	722	-3.99	p<.001*
Średnia długość nieprzerwanej wypowiedzi (w sylabach)	26.21	20.20	2.56	p=.011*
Proporcja wypowiedzi	0.87	0.80	11.38	p<.001*

Różnice w płynności między ST i TT

Liczba i długość cichych i wypełnionych pauz

mniej cichych pauz, dłuższe pauzy ciche i wypełnione w tłumaczeniu (Ahrens 2005; Christodoulides 2013; Collard & Defrancq 2019; Lee 1999; Pöchhacker 1995)

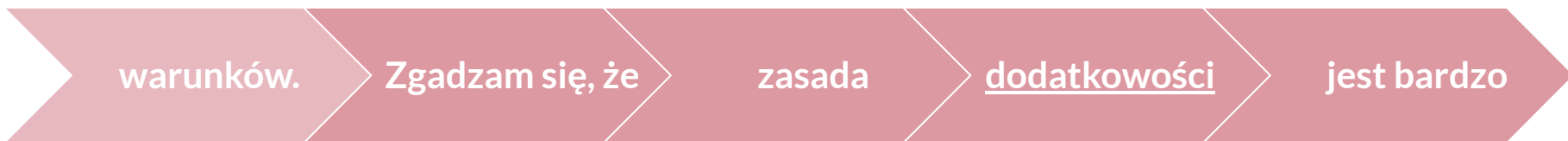
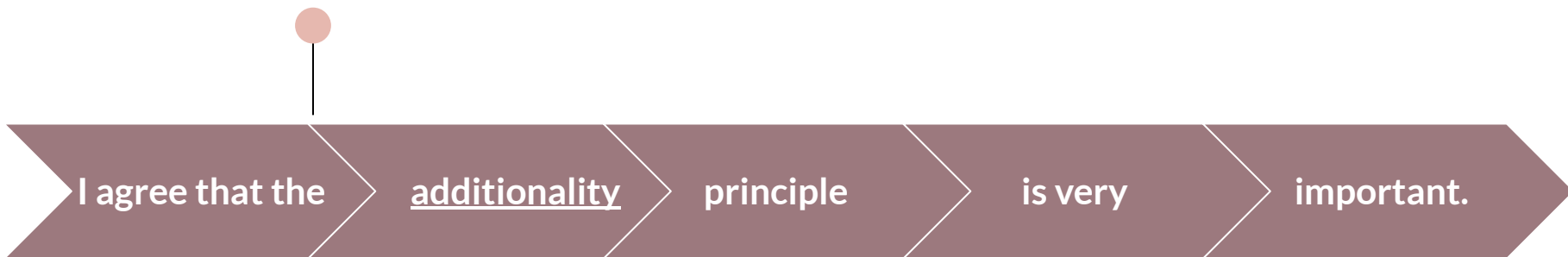
Długość nieprzerwanej wypowiedzi

krótsze fragmenty wypowiedzi w tłumaczeniu, mniej płynna i bardziej fragmentaryczna wypowiedź

Proporcja wypowiedzi

niższa proporcja wypowiedzi w tłumaczeniu

Obciążenie kognitywne: spillover effect



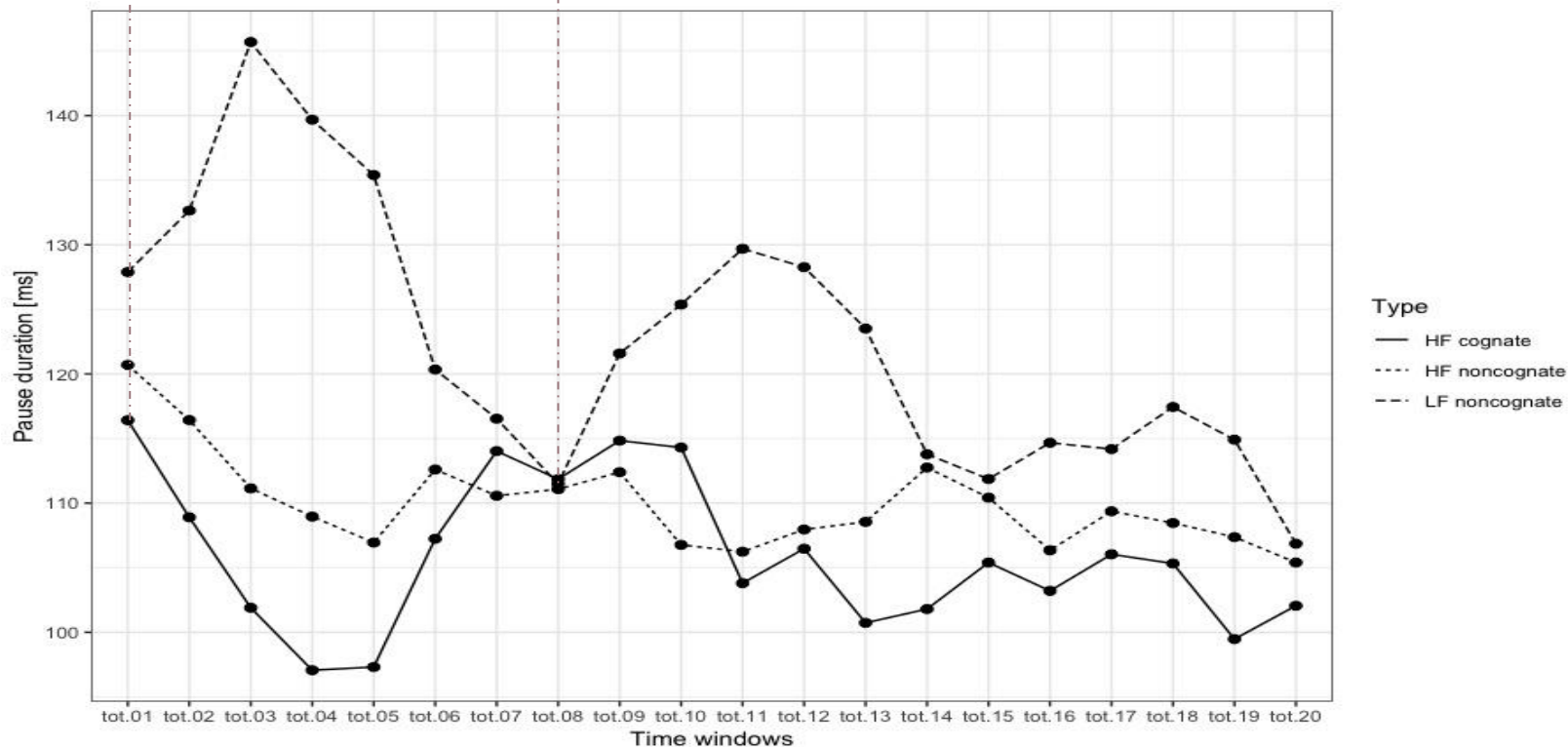
EVS (ear-voice span) ok. 4s

obciążenie importowane
+ bieżące

obciążenie bieżące
+ eksportowane

I agree that the additionality principle is very important.

warunków. Zgadzam się, że zasada dodatkowości jest bardzo



Tłumaczenie kognatów i homonimów

Kognaty i homonimy

million

- aktywacja ekwiwalentu tłumaczeniowego *milion*
- cognate facilitation effect (Dijkstra et. al 2014), model Multilink (Dijkstra et al. 2018)

billion

- inhibicja błędnego odpowiednika *bilion*
 - aktywacja właściwego ekwiwalentu *miliard*
 - Inhibitory Control Model (Green 1998)
-

We know that millions allocated to this program

warunków. Wiemy, że milliony przeznaczone na ten

We know that billions allocated to this program

warunków. Wiemy, że miliardy przeznaczone na

Dalsze badania

1. Identyfikacja struktur gramatycznych charakterystycznych dla polskiego *interpretese*
 2. Analiza jakości tłumaczenia
 3. Porównanie tłumaczenia na angielski L1-L2 vs L2-L1
 4. Korelacje z danymi o pamięci roboczej
- ...



POLISH INTERPRETING CORPUS

Polish Interpreting Corpus:

jak korpus może pomóc w badaniach
nad procesem tłumaczenia
symultanicznego

Agnieszka Chmiel



UNIWERSYTET
IM. ADAMA MICKIEWICZA
W POZNANIU

Danijel Koržinek



POLSKO-JAPOŃSKA AKADEMIA
TECHNIK KOMPUTEROWYCH