Using language corpora and wordnets in the translator's work: pros and cons

Łukasz Grabowski
University of Opole (Poland)

Translating Europe Workshop
23.05.2018, Warsaw

Plan of my talk

- Introduction
  - Language corpora
  - Wordnets
- Using language corpora in the translator's work
  - Parallel corpora
  - Monolingual reference corpora
- Using wordnets in the translator's work
  - Polish wordnet (*Słowosieć*) mapped onto Princeton WordNet of English
- Conclusions

What are language corpora?

- Custom-designed **collections of authentic linguistic data (texts)**, either **written texts** or **transcriptions of recorded spoken data**
- Compiled according to a number of criteria
    - Size
    - Text availability / copyright
    - Representativeness
    - Balance
    - Computer-readable format

Language corpora for translators

- Language corpora specifically designed for translators may include**:**
  - **originals and translations**
    - **Parallel corpora**
      - proved to be of great value for translators (Google Translate, TAUS Data Association, Linguee, Reverso Context etc. are all built on parallel corpora)
      - **Professional translators are more familiar with parallel search** (or parallel concordancing) **feature of translation memory systems** (TMs)
  - **non-translations** (i.e. native texts written originally in a given language)
    - **monolingual reference corpora**
  - **texts in different languages**
    - **originals and translations in the same language**, or
    - only **native texts originally produced in different languages**)
      - **bilingual and monolingual comparable corpora**
      - primarily used for research purposes (Descriptive Translation Studies)

Why should translators bother with language corpora?

- **Corpus as a source of linguistic information** helps to answer various questions related to how we use language
  - across genres, text types and registers
  - **adherence to discourse norms and conventions of text production in the target language** and culture is essential in translation
    - this information is implicitly recorded in texts collected in language corpora
  - particularly important when translators deal with **creative and stylistically varied texts** (rather than cliched and highly-repetitive ones)

- Corpora enable the **translators to go beyond their linguistic intuition**
  - **linguistic preferences of translators**, in particular those rendering texts into a target language which is not their native one, **may not always coincide with the native speakers' preferences**

- Corpora may help **activate the translators' memory** of relevant contexts of use of words and expressions (Pęzik 2017)
  - Lexical priming theory (Hoey 2005)

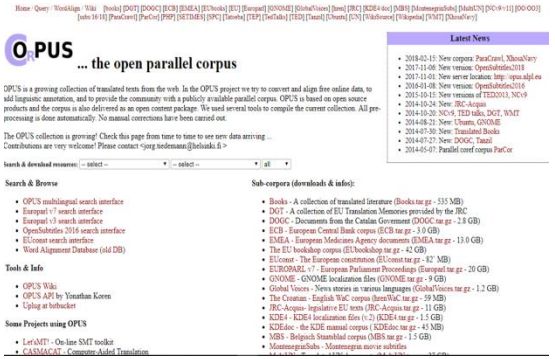
Monolingual reference corpora
Examples

- *British National Corpus*
- *Corpus of Contemporary American English*
- *Narodowy Korpus Języka Polskiego*
- *Национальный корпус русского языка*
- *English Web 2015 (enTenTen15)*
- *Polish Web 2012 (plTenTen12)*
- *Russian Web 2011 (ruTenTen11)*

- Monco (http://monitorcorpus.com/)
  - a collection of English texts updated on a regular basis (circa 11 billion words, as of 9 May 2018)
- Frazeo (http://frazeo.pl)
  - an innovative search engine of press articles published on Polish news portals

OPUS Project
Open Source Parallel Corpus

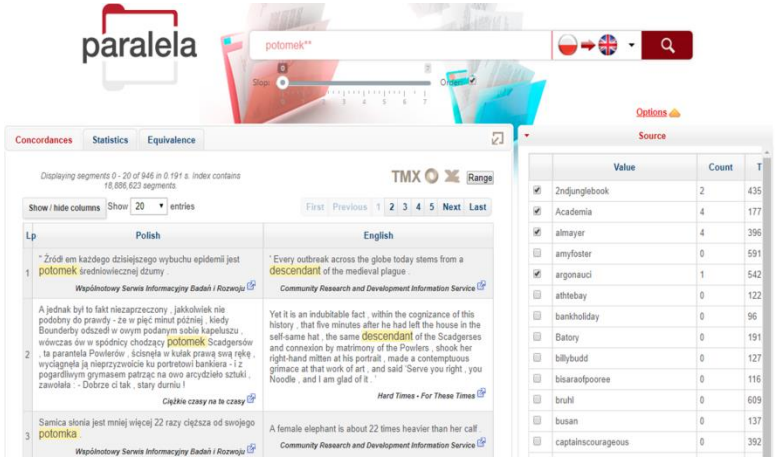
Tiedemann (2009), http://opus.nlpl.eu/



- Collection of translated texts from the web
- Free online data
  - Based on open-source products
  - Delivered as an open content package
    - **Downloadable in many formats, including TMX**
- The corpus contains medical texts, legal texts, software documentation, subtitles etc.

Paralela

(Pęzik 2016)

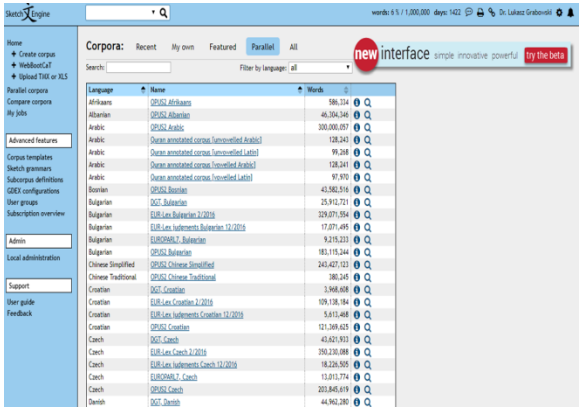


- **An English-Polish and Polish-English parallel corpus**
  - It includes **262 million words in 10,877,000 translation segments** found predominantly in legal texts
    - EU legislation, proceedings of the European Parliament etc.), press releases, medical texts (provided by the EMA) as well as film subtitles
- Translation segments are aligned at the sentence level (Pęzik 2016: 70), with **5.3% of the segments aligned manually**.

Everything in one place

SketchEngine (Kilgarriff et al. 2014)

https://www.sketchengine.co.uk/



- A software platform for text analysis and text mining applications
- Access to **aligned parallel corpora** of various text types and genres
- Access to many **large monolingual corpora** with texts extracted from the web
  - **Term extraction** (from subject-specific texts uploaded by users) and **bilingual term extraction** (performed on TMs uploaded by users)
- A possibility to create **subject-specific corpora**
  - By **uploading materials provided by users**
  - By **providing keywords and extracting relevant texts automatically from the web**

Example 1

- **Multi-word terms** (nominalizations) extracted from a custom-designed corpus of **car reviews** (Top Gear)
  - 100 texts (60,372 words)

Multi-word		Score	F	RefF
inside layout	W	1,385.85	100	0
all-wheel drive	W	230.43	17	3
hot hatch	W	206.09	16	9
driving position	W	167.12	15	28
four-wheel drive	W	155.95	17	58
six-speed manual	W	153.33	11	0
city car	W	149.45	11	3
cruise control	W	139.78	11	11
boot space	W	131.34	10	7
climate control	W	128.08	10	10
turbo engine	W	119.31	9	6
automatic gearbox	W	117.61	10	21
front-wheel drive	W	116.33	9	9
6-litre diesel	W	111.79	8	0
turbo petrol	W	111.79	8	0
petrol engine	W	108.42	12	61
fuel economy	W	104.16	12	68
steering wheel	W	102.21	19	178
body roll	W	101.90	8	11
wind noise	W	100.26	8	13
infotainment system	W	97.94	7	0
4-litre petrol	W	97.16	7	1
driving experience	W	97.12	8	17
engine range	W	96.30	7	2
manual gearbox	W	93.81	7	5
2-litre diesel	W	84.09	6	0
infotainment screen	W	84.09	6	0
trim level	W	83.42	6	1
material quality	W	83.42	6	1
centre console	W	79.86	6	6
base model	W	79.86	6	6
torsion beam	W	70.24	5	0
km co2	W	70.24	5	0
adaptive cruise	W	70.24	5	0
radar cruise	W	70.24	5	0
three-cylinder engine	W	70.24	5	0
boot floor	W	70.24	5	0
rear headroom	W	70.24	5	0
plug-in hybrid	W	70.24	5	0
blind-spot warning	W	70.24	5	0
head-up display	W	69.68	5	1
front end	W	69.26	10	114
5-litre diesel	W	69.07	5	2
six-speed gearbox	W	67.87	5	4
company car	W	67.49	11	143
ride quality	W	67.28	5	5
standard car	W	66.14	5	7
company car tax	W	65.59	5	8
sweet spot	W	64.50	5	10
list price	W	63.85	7	60
rear axle	W	58.34	5	23
old car	W	58.12	7	77
electric power	W	57.87	6	51
cross-traffic alert	W	56.39	4	0
eight-speed auto	W	56.39	4	0
air con	W	56.39	4	0
safety kit	W	56.39	4	0
collision mitigation	W	56.39	4	0
three-cylinder petrol	W	56.39	4	0
electric power steering	W	56.39	4	0
rear suspension	W	56.24	5	28
standard kit	W	55.95	4	1
supple ride	W	55.95	4	1
driving environment	W	55.45	4	2
5-litre turbo	W	55.45	4	2
handling balance	W	54.96	4	3
rear-wheel drive	W	54.96	4	3
6-litre engine	W	54.96	4	3
active safety	W	54.49	4	4
five-year warranty	W	54.49	4	4
rear legroom	W	54.02	4	5
low speed	W	53.95	5	34
standard equipment	W	51.50	5	41
electric motor	W	51.50	5	41
turning circle	W	50.17	4	14
drive system	W	50.17	4	14
car tax	W	49.64	6	78
big car	W	49.22	5	48
automatic transmission	W	45.81	4	26
big news	W	45.15	4	28
small car	W	44.51	5	65
diesel engine	W	43.89	6	103
front passenger	W	43.61	4	33
driver assist	W	42.55	3	0
reversing camera	W	42.55	3	0
practical car	W	42.55	3	0
dual-zone climate	W	42.55	3	0
instrument cluster	W	42.55	3	0
adaptive cruise control	W	42.55	3	0
dual-zone climate control	W	42.55	3	0
radar cruise control	W	42.55	3	0
small crossover	W	42.55	3	0
lane departure	W	42.55	3	0
autonomous emergency	W	42.55	3	0
five-door hatch	W	42.55	3	0
central infotainment	W	42.55	3	0
crossover market	W	42.55	3	0
seven-seat mpv	W	42.55	3	0
5-litre petrol	W	42.55	3	0
supple suspension	W	42.55	3	0

Example 2

- **Bilingual wordsketches**
  - **matched headwords and collocations** extracted from parallel corpora
    - *Declaration/deklaracja* preceded by modifiers in EUROPARL7, English and Polish

declaration *(noun)* EUROPARL7, English freq = 5,532 (91.07 per million)

deklaracja *(noun)* EUROPARL7, Polish freq = 1,620 (106.77 per million)

Use another candidate translation: pisemny oświadczenie patrzeć niepodległość wpisać protokół składać protokół rejestr
Click on collocates to access reciprocal bilingual search or find translated collocations

modifier		52.62
written	236	10.46
written declaration		
Written	76	9.62
: see Minutes Written declarations included in the		
unilateral	77	9.16
unilateral declaration of independence		
joint	258	9.08
a joint declaration		
solemn	53	9.08
solemn declarations		
Lapsed	45	8.96
: see Minutes Lapsed written declarations : see Minutes		
Laeken	35	8.22
the Laeken declaration		
mere	36	8.01
mere declarations of		
grand	19	7.55
the grand declarations		
final	79	7.50
the final declaration		
open	66	7.07
the open declaration		
formal	23	7.06
a formal declaration		
Doha	17	6.98
of the Doha declaration		
non-binding	12	6.97
non-binding declarations		
fine	19	6.85
fine declarations of		
ministerial	12	6.62
the ministerial declaration		
secret	13	6.55
secret declarations		

a modifier		68.58
powszechny	99	10.27
Powszechnej Deklaracji Praw Człowieka		
paryski	21	9.18
deklaracji paryskiej z		
jednostronny	26	9.12
jednostronnej deklaracji niepodległości Kosowa		
pisemny	47	9.09
pisemną deklarację		
gołosłowny	13	8.56
gołosłowne deklaracje		
berliński	11	8.16
deklaracji berlińskiej		
końcowy	14	7.63
deklaracji końcowej		
pusty	8	7.56
krajowy	58	7.47
krajowych deklaracji w sprawie zarządzania		
manhattański	6	7.45
wspólny	96	7.42
wspólnej deklaracji		
praski	6	7.42
jednoznaczny	8	7.29
otwarty	12	7.25
" otwartej deklaracji ”		
samozwańczy	5	7.16
uroczysty	5	7.07
boloński	5	7.07
polityczny	60	6.85
fałszywy	5	6.82
dauhański	5	6.79
jasny	14	6.78
milenijny	6	6.78
wyraźny	11	6.66
ministerialny	4	6.66

  - We can see which collocations for English *declaration* do not have their Polish equivalent with *deklaracja*

Example 2.1

- If we click on collocates, we obtain access to translated collocations

written *(adjective)* Alternative PoS: write verb (freq: 4,279)
EUROPARL7, English freq = 2,431 (40.02 per million)

pisemny *(adjective)* EUROPARL7, Polish freq = 1,700 (112.05 per million)

Use another candidate translation: 142 oświadczenie regulamin 149 art wyjaśnienie) (patrzeć
Click on collocates to access reciprocal bilingual search or find translated collocations

modifies		98.03
statement	1,041	10.10
. Written statements (Rule 149		
explanation	154	9.23
. Written explanations of vote in		
declaration	236	9.17
written declaration		
answer	171	8.25
will receive written answers		
reply	74	8.09
a written reply		
STATEMENT	11	7.23
. WRITTEN STATEMENT (RULE 120		
press	33	7.01
the written press ,		
Statement	7	6.48
consent	11	6.24
written consent		
questionnaire	6	6.15
confirmation	7	6.05
translation	8	5.90
question	150	5.71
a written question		
Declaration	9	5.63
constitution	12	5.53
written constitution		
notification	5	5.52
exam	3	5.29
version	9	5.11
assurance	5	4.95
record	7	4.86
submission	3	4.82
summary	3	4.79
authorisation	4	4.71
form	27	4.68

modifies		99.76
oświadczenie	1,285	12.95
. Oświadczenia pisemne (art .		
wyjaśnienie	123	10.21
. Pisemne wyjaśnienia dotyczące sposobu głosowania		
deklaracja	47	8.79
pisemną deklarację		
tłumaczenie	19	8.26
tłumaczeń pisemnych		
zapytanie	9	7.32
odpowiedź	37	7.20
pisemnej odpowiedzi		
pytanie	37	6.60
pytanie pisemne		
forma	14	6.35
w formie pisemnej		
wersja	6	6.00
kwestionariusz	3	5.82
potwierdzenie	3	5.70
zapis	4	5.62
zgoda	4	5.15
opinia	4	4.14
dokument	3	3.71
poparcie	3	3.68
rezolucja	4	3.62
wniosek	7	3.53
procedura	3	3.51
sprawozdanie	5	2.30

and/or		6.50
verbal	6	8.96
oral	29	7.89
written and oral		
express	4	7.86
prior	3	6.22
audiovisual	3	6.08
formal	4	5.50
detailed	4	4.70
parliamentary	7	4.58
original	3	4.49
annual	3	3.84
comprehensive	3	3.83
urgent	3	3.71
s	3	2.36

When do translators typically use language corpora? Or should use?

- **Monolingual reference corpora** (e.g. *Narodowy Korpus Języka Polskiego*)
  - When we are worried that **the translation does not fit the norms and conventions (grammatical, stylistic etc.) of the target language** (Polish language)
    - Excessive **lexical or syntactic calques** and **idiosyncratic collocations** (*massage baths – wanny masażowe, computer key – guzik komputerowy* etc.), **overuse or underuse of certain grammatical structures** (e.g. passive voice in Polish) or **prefabricated formulas** etc.

- **Parallel corpora**
  - When we want to see **how translators dealt with similar translational problems in the past** (Zanettin 2003), and
    - to verify translational choices / equivalents (textual)
    - to verify dictionary equivalents

Deleveraging – delewarowanie?

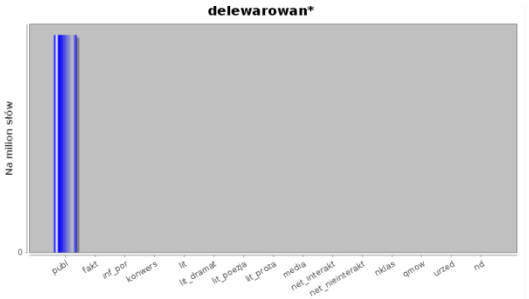
- *Most notably, in the wake of the Lehman collapse, the pattern of **deleveraging** and forced sales has been so intense that traditional price relationships have completely broken down, sending trading models haywire.*
- *W szczególności, po upadku Lehmanów, **delewarowanie** i wymuszone sprzedaże tak się nasiliły, że doszło do całkowitego załamania się tradycyjnych stosunków cenowych i rozchwiania modeli prowadzenia interesów.*
  - Financial world is stumbling blindly through a cognitive fog, Gillian Tett, published 6 March 2009. At: http://www.ft.com/cms/s/0/da26af7c-09ee-11de-add8-0000779fd2ac.html
  - Świat finansów potyka się i błądzi. At: http://ft.onet.pl

Deleveraging – delewarowanie?

- Data from Paralela
  - 12 occurrences of *deleveraging* (10 in RAPID subcorpus (EU Press Releases), 1 in EPP proceedings, 1 in NBP subcorpus)
  - Polish equivalents*:*
    - *oddłużanie*
    - *ograniczanie zadłużenia*
    - *pozbywanie się pasywów*
    - *eliminacja dźwigni finansowej*
    - ***delewarowanie (5 occurrences)*** ✓
    - *zrównoważanie (budżetów domowych) - deleveraging (of households ' balance sheets)*

delewarowanie

- Data from NKJP
  - 10 occurrences of *delewarowanie*

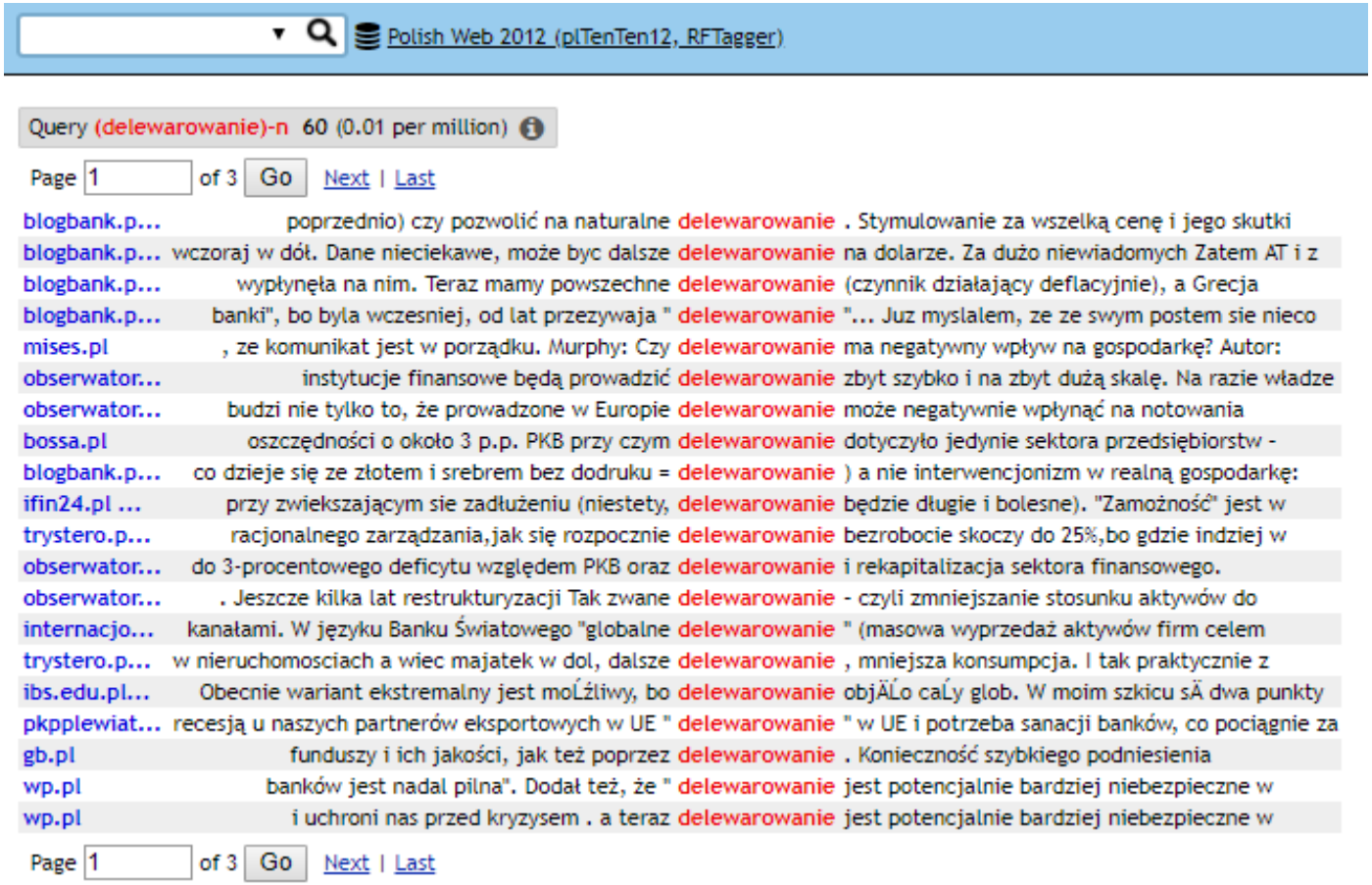


Przeszukiwany zbiór zawiera 240,192,461 słów. Znaleziono **10** akapitów pasujących do zapytania w **0.003s.** Bieżąca strona zawiera 13 przykłady z 9 różnych tekstów.

1.	gdy nadmuchany balon pękł, rozpoczął się odwrotny proces -	**delewarowania**	- czyli zmniejszania się rynku aktywów finansowych, głównie	Polityka	
2.	minionych kilku lat prowadzą dziś intensywny proces	**delewarowania**	, czyli spisywania na straty części aktywów, przy	Polityka	
3.	przebiegało w trzech fazach: fazie stabilizacji, fazie	**delewarowania**	oraz długotrwałej fazie zredukowanego zadłużenia i	Gazeta Ubezpieczeni...	
4.	wzrost produktu krajowego brutto (PKB) w miarę postępującego	**delewarowania**	portfeli konsumentów, wzrostu bezrobocia, stagnacji cen	Gazeta Ubezpieczeni...	
5.	starań rządów i banków centralnych, biorąc pod uwagę procesy	**delewarowania**	w bankach, firmach i gospodarstwach domowych, nie ma zbyt	Gazeta Ubezpieczeni...	
6.	finansowej z sektora prywatnego do publicznego. Skala	**delewarowania**	potwierdza, że potencjalny wzrost w większości krajów	Gazeta Ubezpieczeni...	
7.	pełni dojrzeć do wzrostów. Sądzimy zatem, że trwający proces	**delewarowania**	i wysoki stopień wzajemnych zależności ekonomicznych w skali	Gazeta Ubezpieczeni...	
8.	azjatyckich w połączeniu z trwającym w tej chwili procesem	**delewarowania**	w USA mogą w przyszłych latach wyprowadzić Azję na ścieżkę	Gazeta Ubezpieczeni...	
9.	minęła pierwsza panika, w wieżowcach na Manhattanie ruszyło	**delewarowanie**	. W czasach hossy ulubionym narzędziem bankierów z Wall	Polityka	
10.	wskutek utraty zaufania do instytucji finansowych, z drugiej	**delewarowanie**	, czyli pozbywanie się nadmiernego zadłużenia przez banki i	Polityka	
11.	ale póki trwa, podgrzewa koniunkturę w gospodarce.	**Delewarowanie**	jest niezbędne, ale procesowi temu towarzyszyć muszą	Polityka	
12.	rządy na ratowanie sektora bankowego bilionów dolarów - bo	**delewarowanie**	to proces, którego skala nie idzie w biliony, ale grube	Polityka	
13.	o wysokiej rentowności i kredytów bankowych. Błyskawiczne	**delewarowanie**	systemu finansowego także miało swój negatywny wpływ na	Gazeta Ubezpieczeni...	

Delewarowanie

- Data from plTenTen12
  - 60 occurrences of *delewarowanie*



Polish Web 2012 (plTenTen12, RFTagger)

Query (delewarowanie)-n 60 (0.01 per million)

Page 1 of 3 Go Next | Last

blogbank.p...	poprzednio) czy pozwolić na naturalne	delewarowanie	. Stymulowanie za wszelką cenę i jego skutki
blogbank.p...	wczoraj w dół. Dane nieciekawe, może byc dalsze	delewarowanie	na dolarze. Za dużo niewiadomych Zatem AT i z
blogbank.p...	wypłynęła na nim. Teraz mamy powszechne	delewarowanie	(czynnik działający deflacyjnie), a Grecja
blogbank.p...	banki", bo byla wczesniej, od lat przezywaja "	delewarowanie	"... Juz myslalem, ze ze swym postem sie nieco
mises.pl	, ze komunikat jest w porządku. Murphy: Czy	delewarowanie	ma negatywny wpływ na gospodarkę? Autor:
obserwator...	instytucje finansowe będą prowadzić	delewarowanie	zbyt szybko i na zbyt dużą skalę. Na razie władze
obserwator...	budzi nie tylko to, że prowadzone w Europie	delewarowanie	może negatywnie wpłynąć na notowania
bossa.pl	oszczędności o około 3 p.p. PKB przy czym	delewarowanie	dotyczyło jedynie sektora przedsiębiorstw -
blogbank.p...	co dzieje się ze złotem i srebrem bez dodruku =	delewarowanie	) a nie interwencjonizm w realną gospodarkę:
ifin24.pl ...	przy zwiekszającym sie zadłużeniu (niestety,	delewarowanie	będzie długie i bolesne). "Zamożność" jest w
trystero.p...	racjonalnego zarządzania,jak się rozpocznie	delewarowanie	bezrobocie skoczy do 25%,bo gdzie indziej w
obserwator...	do 3-procentowego deficytu względem PKB oraz	delewarowanie	i rekapitalizacja sektora finansowego.
obserwator...	. Jeszcze kilka lat restrukturyzacji Tak zwane	delewarowanie	- czyli zmniejszanie stosunku aktywów do
internacjo...	kanałami. W języku Banku Światowego "globalne	delewarowanie	" (masowa wyprzedaż aktywów firm celem
trystero.p...	w nieruchomosciach a wiec majatek w dol, dalsze	delewarowanie	, mniejsza konsumpcja. I tak praktycznie z
ibs.edu.pl...	Obecnie wariant ekstremalny jest moĹźliwy, bo	delewarowanie	objÄĹo caĹy glob. W moim szkicu sÄ dwa punkty
pkpplewiat...	recesją u naszych partnerów eksportowych w UE "	delewarowanie	" w UE i potrzeba sanacji banków, co pociągnie za
gb.pl	funduszy i ich jakości, jak też poprzez	delewarowanie	. Konieczność szybkiego podniesienia
wp.pl	banków jest nadal pilna". Dodał też, że "	delewarowanie	jest potencjalnie bardziej niebezpieczne w
wp.pl	i uchroni nas przed kryzysem . a teraz	delewarowanie	jest potencjalnie bardziej niebezpieczne w

Page 1 of 3 Go Next | Last

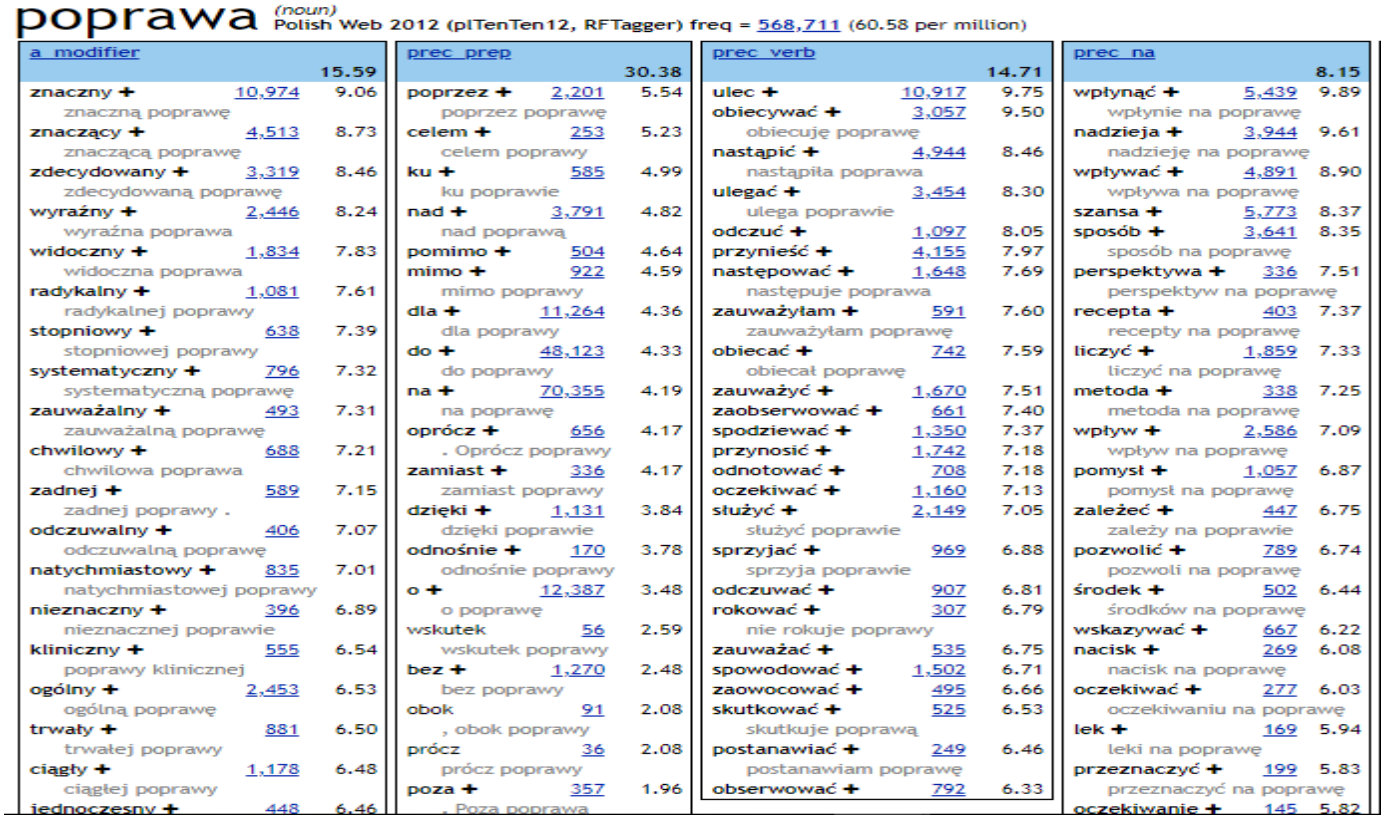
Sprowokować poprawę?

- Example from Paralela (JRC-Acquis sub-corpus)
  - *Another important issue is then discussed — the forced sale of mortgaged property . The GP notes that this area is characterised by a large number of procedures , timeframes and costs and that this would hinder cross-border lending activity. It therefore suggests a gradual approach to* ***encourage improvements*** *in forced sales procedures: (…)*
  - *Inną ważną kwestią omówioną w dalszej kolejności jest wymuszona sprzedaż nieruchomości obarczonej hipoteką . ZK zaznacza , że obszar ten charakteryzuje duża liczba procedur , okresów i kosztów i stanowiłoby to przeszkodę dla transgranicznej działalności kredytowej . Sugeruje zatem zastosowanie stopniowego podejścia do* ***sprowokowania poprawy*** *w zakresie procedur wymuszonej sprzedaży*

Sprowokować poprawę

- Data from NKJP
  - No occurrences of *sprowokować poprawę*
- Data from Frazeo
  - No occurrences of *sprowokować poprawę*
- Data from plTenTen12
  - 2 occurrences of *sprowokować poprawę* (in 7.7 billion words)
    - Jakość mediów – tych drukowanych i tych elektronicznych – jego zdaniem, dramatycznie spada. Na dodatek, nie widać żadnego mechanizmu w mediach prywatnych, który mógłby ***sprowokować poprawę*** sytuacji. (Source: http://www.dziennikarzerp.pl/2010/12/kulawa-debata-publiczna/)
    - Dodam jeszcze, że zamiast ubolewać, co jest nie tak, trzeba było samemu ***sprowokować poprawę*** - ale nie przez narzekanie, tylko przez rzucenie konkretnymi propozycjami- co i jak poprawić. (Source: http://www.forum.fantazyzone.pl/printview.htm?t=51&start=0&sid=63ee71cac16f6f8ccdd3d53d34eb6104)

- Data from plTenTen12
  - *Przynieść poprawę, sprzyjać poprawie, (s)powodować poprawę, zaowocować poprawą, skutkować poprawą*



poprawa *(noun)*
Polish Web 2012 (plTenTen12, RFTagger) freq = 568,711 (60.58 per million)

a_modifier		15.59
znaczny +	10,974	9.06
znaczną poprawę		
znaczący +	4,513	8.73
znaczącą poprawę		
zdecydowany +	3,319	8.46
zdecydowaną poprawę		
wyraźny +	2,446	8.24
wyraźna poprawa		
widoczny +	1,834	7.83
widoczna poprawa		
radykalny +	1,081	7.61
radykalnej poprawy		
stopniowy +	638	7.39
stopniowej poprawy		
systematyczny +	796	7.32
systematyczną poprawę		
zauważalny +	493	7.31
zauważalną poprawę		
chwilowy +	688	7.21
chwilowa poprawa		
zadnej +	589	7.15
zadnej poprawy .		
odczuwalny +	406	7.07
odczuwalną poprawę		
natychmiastowy +	835	7.01
natychmiastowej poprawy		
nieznaczny +	396	6.89
nieznacznej poprawie		
kliniczny +	555	6.54
poprawy klinicznej		
ogólny +	2,453	6.53
ogólną poprawę		
trwały +	881	6.50
trwałej poprawy		
ciągły +	1,178	6.48
ciągłej poprawy		
jednoczesny +	448	6.46

prec_prep		30.38
poprzez +	2,201	5.54
poprzez poprawę		
celem +	253	5.23
celem poprawy		
ku +	585	4.99
ku poprawie		
nad +	3,791	4.82
nad poprawą		
pomimo +	504	4.64
mimo +	922	4.59
mimo poprawy		
dla +	11,264	4.36
dla poprawy		
do +	48,123	4.33
do poprawy		
na +	70,355	4.19
na poprawę		
oprócz +	656	4.17
. Oprócz poprawy		
zamiast +	336	4.17
zamiast poprawy		
dzięki +	1,131	3.84
dzięki poprawie		
odnośnie +	170	3.78
odnośnie poprawy		
o +	12,387	3.48
o poprawę		
wskutek	56	2.59
wskutek poprawy		
bez +	1,270	2.48
bez poprawy		
obok	91	2.08
, obok poprawy		
prócz	36	2.08
prócz poprawy		
poza +	357	1.96
. Poza poprawa		

prec_verb		14.71
ulec +	10,917	9.75
obiecywać +	3,057	9.50
obiecuję poprawę		
nastąpić +	4,944	8.46
nastąpiła poprawa		
ulegać +	3,454	8.30
ulega poprawie		
odczuć +	1,097	8.05
przynieść +	4,155	7.97
następować +	1,648	7.69
następuje poprawa		
zauważyłam +	591	7.60
zauważyłam poprawę		
obiecać +	742	7.59
obiecał poprawę		
zauważyć +	1,670	7.51
zaobserwować +	661	7.40
spodziewać +	1,350	7.37
przynosić +	1,742	7.18
odnotować +	708	7.18
oczekiwać +	1,160	7.13
służyć +	2,149	7.05
służyć poprawie		
sprzyjać +	969	6.88
sprzyja poprawie		
odczuwać +	907	6.81
rokować +	307	6.79
nie rokuje poprawy		
zauważać +	535	6.75
spowodować +	1,502	6.71
zaowocować +	495	6.66
skutkować +	525	6.53
skutkuje poprawą		
postanawiać +	249	6.46
postanawiam poprawę		
obserwować +	792	6.33

prec_na		8.15
wpłynąć +	5,439	9.89
wpłynie na poprawę		
nadzieja +	3,944	9.61
nadzieję na poprawę		
wpływać +	4,891	8.90
wpływa na poprawę		
szansa +	5,773	8.37
sposób +	3,641	8.35
sposób na poprawę		
perspektywa +	336	7.51
perspektyw na poprawę		
recepta +	403	7.37
recepty na poprawę		
liczyć +	1,859	7.33
liczyć na poprawę		
metoda +	338	7.25
metoda na poprawę		
wpływ +	2,586	7.09
wpływ na poprawę		
pomysł +	1,057	6.87
pomysł na poprawę		
zależeć +	447	6.75
zależy na poprawie		
pozwolić +	789	6.74
pozwoli na poprawę		
środek +	502	6.44
środków na poprawę		
wskazywać +	667	6.22
nacisk +	269	6.08
nacisk na poprawę		
oczekiwać +	277	6.03
oczekiwaniu na poprawę		
lek +	169	5.94
leki na poprawę		
przeznaczyć +	199	5.83
przeznaczyć na poprawę		
oczekiwanie +	145	5.82

sprowokować

- Sprowokować *coś*
- Sprowokować *kogoś do czegoś*

has_obj4		4.00
zajście	34	8.46
sprowokował całe zajście		
dyskusja +	225	8.02
zamieszki	16	7.89
reakcja +	126	7.49
bójka	9	7.23
debata	25	7.15
skandal	12	6.98
awantura	22	6.92
kłótnia	11	6.92
konflikt	31	6.77
odwet	7	6.68
wybuch	12	6.63
pojawić	9	6.48
zamieszka	5	6.44
incydent	8	6.40
odzew	7	6.38
interwencja	9	6.34
burda	5	6.28
zadyma	5	6.19
katastrofa	9	6.13
wojna	50	6.10
korekta	15	6.09
lawina	6	6.08
inwazja	5	6.01
refleksja	13	5.98

» the verb *sprowokować ,'provoke sth'* has a negative aura of meaning in Polish

post_do		4.91
kłótnia	66	8.16
się sprowokować do kłótni		
brać	54	8.08
sprowokować do brania		
bójka	34	7.75
sprowokować do bójki .		
pyskówka	15	7.63
reakcja	41	7.58
agresja	28	7.58
sprowokować do agresji .		
zastanowić	34	7.36
sprowokować do zastanowienia się nad		
dyskusja +	325	7.18
sprowokować do dyskusji		
przemyślenie	39	7.13
sprowokować do przemyśleń		
rozprawiać	9	7.01
prawić	9	7.01
myślenie +	120	6.96
sprowokować do myślenia		
odpowiedź	41	6.95
sprowokować do odpowiedzi		
polemika	16	6.93
zabrać	17	6.93
sprowokować do zabrania głosu		
bałakać	8	6.85
awantura	19	6.84
sprowokować do awantury		
odezwać	8	6.84
atak +	116	6.83
sprowokować do ataku .		
zadać	14	6.83
sprowokować do zadania		
wypowiedź	70	6.68
sprowokować do wypowiedzi		
pleść	7	6.66
trajkotać	7	6.66
powiadać	7	6.66
zachowanie	35	6.64

Limitations of language corpora

from the perspective of translators

- **Lack of sense disambiguation**
  - When we translate, we search for sense rather than form
    - *e.g. balkonik* (see later example)
      - framework for walking, part of a buliding, an undergarment?
- **Limited size of some language corpora**
  - **Negative evidence**
    - the fact that a given word or expression does not occur in the corpus does not mean that it is not used
- **If not updated, corpora do not catch up with reality**
  - No occurrences of *fake news, lokowanie produktu, hejtować* in NKJP
- **Loads of data mean problems with their filtering and interpretation**
  - The need to learn corpus functionalities (corpus search engines, query syntax etc.)
  - **Translators are busy people**
    - They may not have time for learning all this
      - Resource allocation for specific roles in the translation project

Wordnets

- Large **relational lexico-semantic databases**
- **Princeton WordNet** (Fellbaum 1998) was built on psycholinguistic principles
  - http://wordnetweb.princeton.edu/perl/webwn
  - **intended to reflect the structure of human lexical memory** (cf. Miller 1998)
    - originated as an experiment on mapping lexical memory of children,
    - gradually evolved into a huge electronic resource covering a large part of the lexical system of English (Fellbaum, 1998).
- Words are grouped in the so-called **synsets**
  - nouns, verbs, adjectives and adverbs
  - **represent lexicalised concepts** (Miller 1998)
  - **(cognitive) sets of synonymous lexical units (LUs)**
    - Synonymy is a **conceptual relation established on the basis of a linguist's intuition and dictionary definitions**
- **Synsets** are linked via semantic relations
  - hyponymy, hypernymy, partial-synonymy etc.

Polish wordnet (plWordNet), Słowosieć

- http://plwordnet.pwr.wroc.pl/wordnet/
- In PlWordNet, **synsets are sets of synonymous LUs that enter the same constitutive relations**
  - e.g. hyponymy, hypernymy, meronymy, antonymy (Piasecki et al. 2009, Maziarz et al. 2012)
- **Each word is defined implicitly with reference to other words**
  - *samochód* 'car' is a kind of *pojazd drogowy* 'road vehicle'; it is a whole consisting of *silnik* 'engine', *spryskiwacz* 'windscreen washer', *podwozie* 'chassis' and so on; its close counterpart is the colloquial *fura* and *bryka* 'wheels, etc.

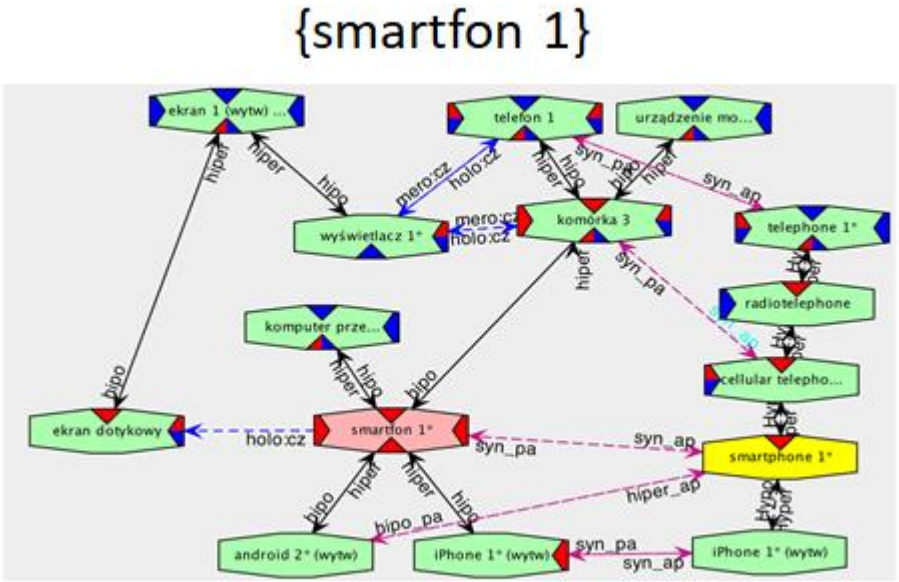
		No. of lemmas	No. of LUs	No. of synsets	Monosemous lemmas	Polysemous lemmas
plWordNet	Verbs	19984	41017	29892	11283	8701
	Nouns	133845	176938	132628	109652	24193
	Adverbs	8010	14040	11260	4692	3318
	Adjectives	29228	54021	46705	16580	12648
	All	191067	286016	220485	142207	48860

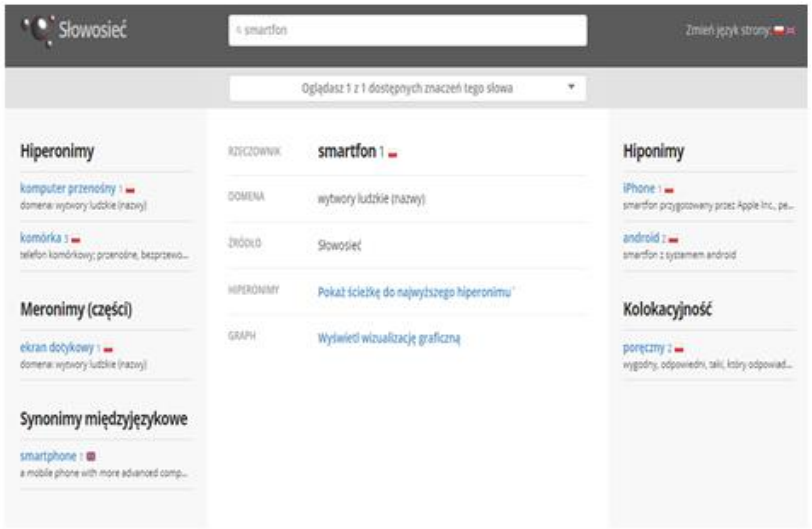
Mapping of wordnets (at synset level)

- **Linking plWordNet and Princeton WordNet synsets** corresponding in **meaning and position in wordnet structure**
  - Currently nouns
- Mapping direction: **plWordNet > Princeton WordNet**
  - **using a set of interlingual relations**
    - **I-synonymy**
    - **I-hyponymy and hypernymy**
    - **I-meronymy and holonymy**
    - **I-partial synonymy**
    - **I-inter register synonymy**
    - **I interparadigmatic synonymy**

Wordnets as bilingual lexicons

- **Wordnets provide a list of senses** of a given word, yet unlike traditional dictionaries, **they also place each specific sense** - in the form of a synset - **in a network of lexico-semantic relations**
- Example





Using wordnets in translation practice

- Using mapped wordnets can be summarized as follows:
  - **Determining the sense of a source language word**
  - **Finding possible translations (candidates)**
  - **Choosing a target language equivalent (a definitive translation)** (Rudnicka & Piasecki 2013)

- Examples

Situation 1

A known word in an unfamiliar context

- *After I finished studying for my courses one weekend, I called my mum. I told her that we were going to have a confirmation service during **chapel**, and I was not looking forward to attending. The topics chosen for **chapel** often seemed negative and left me feeling down.*
  - *Encountering Angels: True Stories of How They Touch Our Lives Every Day*, by Judith MacNutt (2016), Chosen Books*
    - Source: https://books.google.pl/books?id=mgI_CgAAQBAJ&pg=PT23&dq=%22during+chapel%22&hl=pl&sa=X&ved=0ahUKEwi0iu6kpqrbAhUD2SwKHQYODd4Q6AEIMDAB#v=onepage&q=%22during%20chapel%22&f=false

* Thanks to E. Rudnicka and T. Piotrowski (2017) for suggesting this example

chapel

- Synsets in Princeton WordNet
  - **{chapel 1}** (artefact)
    - a place of worship that has its own altar
  - **{chapel 2;** chapel service 1**}** (act)
    - a service conducted in a place of worship that has its own altar

- Relevant synset in Princeton Wordnet
  - {**chapel service 1; chapel 2**} (act)
    - a service conducted in a place of worship that has its own altar;
      - he was late for chapel
    - Hypernymy to {service 3, religious service 1, divine service 1}
    - I-hypernymy {**nabożeństwo 1**} ✓
      - forma modlitewnego zgromadzenia wiernych danego wyznania

Situation 1.2

We don’t know the equivalent

*At **Low Mass** with a congregation present, or if the church is large, incense may be blessed beforehand by the celebrant. At Missa Cantata, or **Low Mass** with an opening hymn, the **thurible** may be taken in procession.*

- Cooper, I. (2010). Ceremonies of The Young Rite, p. A-3

Source: https://books.google.pl/books?id=hxX9AgAAQBAJ&pg=RA1-PA3&dq=%22low+mass%22+%2B+church&hl=en&sa=X&ved=0ahUKEwjlr5fQ-MPaAhUQmrQKHaNhA544ChDoAQg8MAU#v=onepage&q=%22low%20mass%22%20%2B%20church&f=false

- Relevant synsets in Princeton WordNet
  - **{low mass 1}**
    - a Mass recited without music
  - Hypernymy to {Mass 4}
    - (Roman Catholic Church and Protestant Churches) the celebration of the Eucharist
  - I-synonymy to **{cicha msza 1}** ✓
    - msza, podczas której nie ma muzyki, diakona ani subdiakona, a na ołtarzu palą się dwie świece; inaczej nazywana mszą czytaną
  - {**thurible 1**; censer 1}
    - a container for burning incense (especially one that is swung on a chain in a religious ritual)
  - Hypernymy to {vessel 3}
    - an object used as a container (especially for liquids)
  - I-synonymy to **{kadzielnica 1; trybularz 1}** ✓
    - naczynie kościelne, które służy do spalania kadzidła

Situation 2

We know the meaning of a SL word but we don't know its TL equivalent

*At Low Mass with a congregation present, or if the church is large, incense may be blessed beforehand by the **celebrant**. At Missa Cantata, or Low Mass with an opening hymn, the thurible may be taken in procession.*

- Cooper, I. (2010). Ceremonies of The Young Rite, p. A-3

- Synsets in Princeton Wordnet
  - {**celebrant 1**; celebrater 1; celebrator 1}
    - A person who is celebrating
  - {**celebrant 2**}
    - an officiating priest celebrating the Eucharist

- Relevant synset in Princeton WordNet
  - {celebrant 2}
    - an officiating priest celebrating the Eucharist
  - Hypernymy to {priest 1}
    - a clergyman in Christian churches who has the authority to perform or administer various religious rites
  - l-synonymy to {**celebrans 1; celebrant 1; oficjant 1**} ✓
    - duchowny odprawiający nabożeństwo

Situation 3

We are not sure about the meaning of a SL word and we don't know its TL equivalent

- *Kiedy przechodząc do kuchni pokonuje próg,* ***balkonik****, o który się opiera, wydaje cichy, ale wyraźny stukot.*
  - (Super Express, 2006; PWN_1302900002719)

Synsets in Słowosieć

- {**balkonik 1**} (art) ✓
  - specjalne urządzenie na kółkach, przy pomocy którego osoby starsze, niepełnosprawne i mające problem z chodzeniem, mogą się łatwiej poruszać
- {**balkonik 2**} (loc)
  - Markedness (diminutives) to {balkon 1}
    - element architektoniczny, przestrzeń wystająca poza mury budynku, często ogrodzona balustradą
- {**balkonik 3**; balkonetka 1} (art.)
  - rodzaj biustonosza z miseczkami skrojonymi z trzech części, który podnosi i zaokrągla piersi

- *Kiedy przechodząc do kuchni pokonuje próg, **balkonik**, o który się opiera, wydaje cichy, ale wyraźny stukot*. (Super Express, 2006; PWN_1302900002719)
- Relevant PlWordNet synset
  - {**balkonik 1**} (domain: artefact)
    - specjalne urządzenie na kółkach, przy pomocy którego osoby starsze, niepełnosprawne i mające problem z chodzeniem, mogą się łatwiej poruszać
  - Hypernymy to {chodzik 1} {podpora 1; podpórka 1; podparcie 2}
  - I-synonymy to {**walker 5; Zimmer frame 1; Zimmer 1**}
    - a light enclosing framework (trade name Zimmer) with rubber castors or wheels and handles ✓

Situation 4

A known word in an unfamiliar context

- Skoro jest wilgoć, to i pojawiły się wreszcie moje **kołpaki** (…) Moje dwie gromady, umiejscowione w odległych od siebie miejscach, wysypują się zawsze w październiku. Chociaż podobno powinny to czynić już od maja. Grzyby te trochę mi się kojarzą z czubajkami kaniami- także mają gładki jasnobrązowy wierzchołek kapelusza i odstające łuski – u młodych białawe, u starszych brązowiejące.
  - Source: http://puszcza.net.pl/ciesza-sie-sojki-i-slimaki
    - Czernidłak kołpakowaty

- **Identifying the sense of the source language word**
  - {**kołpak 1**; dekiel 1}
    - osłona felgi samochodowej
      - Kołpak może być wykonany z metalu lub tworzyw sztucznych.
  - {**kołpak 2**}
    - nakrycie głowy pochodzenia tureckiego, popularne w XIV-XVII wieku w Europie Wschodniej; była to czapka bez daszka, obrzeżona futrem wilczym, rysim lub lisim z aksamitną główką, niekiedy przybierała kształt cylindryczny, zakończony spiczasto
      - Tatarskie kołpaki przypominała za to Czapka Monomacha, nakrycie głowy moskiewskich carów, wskazując też na powiązania kulturowe państw ruskich stanowiących lenno Złotej Ordy.
  - {**kołpak 3;** bełdka kołpak 1; sowa 2; kołpaczek 2; czubajka 2} ✓
    - inna nazwa czernidłaka kołpakowatego
  - {**kołpak 4**; dzwoniak 1; delikatka 1; rumieniaczek 1; bedłka 2; rumieniak 2}
    - ludowa nazwa niektórych grzybów z rodzaju dzwonkówek
  - {**kołpak 5**; podbołotuszka; rozeta pomarszczona 1; płachetka 1; bydlarka 1; niemka 2; płachta 2; turek 3}
    - inna nazwa płachetki zwyczajnej - grzyba jadalnego z rodziny zasłonakowatych

- **Finding possible translation candidates**
  - {**kołpak 3;** bełdka kołpak 1; sowa 2; kołpaczek 2; czubajka 2}
    - inna nazwa czernidłaka kołpakowatego
  - Inter-register synonyms
    - {czernidłak kołpakowaty 1}
  - **I-inter-register synonyms**
    - {shaggymane 1; Coprinus comatus 1; shaggy cap 1; shaggymane mushroom 1}
- **Choosing a target language equivalent**
  - {**shaggymane 1**; Coprinus comatus 1; **shaggy cap 1; shaggymane mushroom 1**} ✓

Limitations of wordnets

- **Polish WordNet has been mapped onto Princeton WordNet only partially**, which translates into **limited coverage**
  - and at the level of synsets (abstract units)
- **Translators do not deal with synsets but with lexical units** (form-meaning mappings)
  - Mapping at the level of lexical units (nouns) is in the pipeline (Rudnicka et al. 2017)
    - It considers formal, semantic and **translational criteria**
- **There may be gaps and mismatches between PWN and plWN (**Rudnicka et al. 2016), which are due to:
  - Wordnet-specific structural and methodological differences
  - Specificity of inter-wordnet mapping procedure
  - Systemic differences between English and Polish
  - Differences in terms of lexicalization of concepts
    - often due to cross-cultural differences
- **Wordnets for different languages (other than English) are created using different methods**
  - Transfer, transfer-and-merge (automatic), transfer-and-merge (semi-automatic and corpus-based) - plWN

Conclusions

- Using parallel and reference corpora as well as wordnets is **a realistic scenario in the translator's work**
  - It offers **exposure to authentic linguistic data** ☺
  - It may help **improve the textual fit of translations** ☺
  - But it can be **time-consuming and labour-intensive** and it may require learning new skills related to the use of those resources ☹
- **Language corpora and wordnets may complement (e-)dictionaries, CATs and TMs**
  - Notably when translating creative texts

References (selected)

- Grabowski, Ł. (2018). “Stance bundles in English-to-Polish translation: a corpus-informed study”. *Russian Journal of Linguistics*, 22 (2), 404-422
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubícek, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). “The Sketch Engine: ten years on”. *Lexicography* 1(1): 7–36.
- Krueger, R. (2017). Working with corpora in the translation classroom
- Pęzik, P. (2017). Korpusy referencyjne, korpusy równoległe, ekwiwalencja frazeologiczna. Presentation delivered at CLARIN-PL workshops (3-4 Feb 2017) in Łódź
- Podhajecka, M. (2013). Calques in Polish translations of English press articles: Linguistic innovations or mistranslations?” In B. Lewandowska-Tomaszczyk and M. Thelen (eds.), Translation and Meaning Part 10. Maastricht: Zuyd University of Applied Sciences. 165−179.
- Rudnicka, E. & Piotrowski, T. (2015). Dwujęzyczna Słowosieć – możliwości wykorzystania w pracy tłumacza. Presentation delivered at CLARIN-PL workshops in IBL in Warsaw.
- Rudnicka, E., Witkowski, W. & Grabowski, Ł. (2016). “Towards a methodology for filtering out gaps and mismatches across wordnets: the case of noun synsets in plWordNet and Princeton WordNet”. In: B. Barbu Mititelu, C. Forascu, Ch. Fellbaum, P. Vossen (Eds.), *Proceedings of the Eighth International Global WordNet Conference 2016,* 27-30 Jan 2016, Bucharest, Romania, pp. 344-351
- Rudnicka, E., Bond, F., Grabowski, Ł., Piasecki, M. & Piotrowski, T. (2017). “Towards equivalence links between senses in plWordNet and Princeton WordNet”. *Lodz Papers in Pragmatics*, 13 (1), 3-24.
- Zanettin, F. (2002). Corpora in Transation Practice. In: Proceedings of the LREC Workshop, Language Resources for Translation Work and Research. 2002. 10-14.

Thank you

- for your attention
  - lukasz@uni.opole.pl
  - ResearchGate: LukaszGrabowski2