UWM Olsztyn Acta Polono-Ruthenica XXIII/2, 2018
ISSN 1427-549X

Yury Fedorushkov
Uniwersytet im. Adama Mickiewicza w Poznaniu

Prolegomena do tagowania frazemów w równoległym korpusie rosyjsko-polskim (literatura piękna) w aspekcie przekładoznawczym[1]

Добрый человек из доброго сокровища сердца своего выносит доброе, а злой человек из злого сокровища сердца своего выносит злое, ибо от избытка сердца говорят уста его.|Dobry człowiek z dobrego skarbca swego serca wydobywa dobro, a zły człowiek ze złego skarbca wydobywa zło. Bo z obfitości serca mówią jego usta.[2]

Wstęp

Nawet pojedynczy akt tłumaczenia przypomina podróż. Podróż tłumaczonego obiektu często ma ścieżkę zawiłą, rozwidlającą się dodatkowo na mikrodecyzje tłumacza; często algorytm powstania pary przekładowej jest skomplikowany, przechodzący w skrajność[3]. Podróż ze skutecznym dotarciem do celu możliwa jest w przypadku wykorzystania systemu sztucznej inteligencji (SI). Podróż ta jest…

[1] Pragnę podziękować Panu dr. Filipowi Gralińskiemu za konsultacje dotyczące praktycznych aspektów korpusologii, zastosowania systemu uczącego się (sztucznej inteligencji), za pomoc przy urównolegleniu zdań rosyjskich i polskich (zastosowanie oprogramowania *bleu-champ* oraz *Moses*).

[2] Cytat z korpusu równoległego pt. *Polsko-rosyjski i rosyjsko-polski korpus równoległy* [Zob. (online) http://pol-ros.polon.uw.edu.pl]. W korpusie wykorzystano tekst *Biblii Tysiąclecia* [Łk 6,45] oraz tzw. przekład synodalny *Biblii* [Лк 6: 45].

[3] Mówiąc o „skrajnym" tłumaczeniu obiektu języka A na obiekt w języku B, mamy na myśli m.in. taką wyjątkową sytuację, kiedy właśnie nie adaptujemy obiektu języka A w języku B, tylko tworzymy obiekt nowy. Jesteśmy wówczas narażeni na tzw. *superbłędy*, tj. błędy świadomie zaadaptowane i w istocie swojej w ostatecznym rozrachunku będące najlepszymi strategiami tłumaczenia – np. w tłumaczeniu limeryków o treści absurdalnej, w której desygnat X w systemie A zastępuje się desygnatem Y w systemie B, tj. zupełnie innym: zmienia się w istocie swojej sens translanda na korzyść z r o z u m i e n i a przez użytkowników systemu B samej strategii illokucyjnej (np. opartej na farsie, karykaturalności sytuacji). Brak desygnatu w języku B wymaga zastąpienia go desygnatem, który spełnia podobną funkcję, ma podobną wagę kulturalną, rolę w sytuacji itp. – np. *flute/лютня*, *jig/частушка* – por. niżej:

There was a young lady of Bute	Жила-была молодая леди в Бьюте,
Who played on a silver-gilt **flute**;	Игравшая на позолоченной **лютне**.
She played several **jigs**	Играла она **частушки**
To her uncle's white pigs (…)	Дядиной поросюшке (…)

Zob. artykuł o limerykach z przykładami tłumaczeń [Ражева 2006, 330].

„daleka". Plusem jest jednak to, że w rachubę wchodzą większe masy tekstowe. Czy tłumaczenie automatyczne „zabiera chleb" tłumaczom? Nie. Wykwalifikowany tłumacz lepiej i „po ludzku" pomoże pokonać tę ścieżkę, tj. uwzględni także naiwny bądź też mentalnościowo-kulturalny obraz świata: *снег идет* (dosł. po pol. *śnieg idzie*) – *śnieg pada* (dosł. po ros. *снег падает*). Jednak po rosyjsku *падает снег* nie jest błędem (por. tytuł filmu animowanego *Падал прошлогодний снег*[4]); *po polsku jednak idzie śnieg* oznacza zupełnie co innego: to, że śnieg się zbliża, nadchodzi (*idzie śnieg*, *idzie zima*). SI nie jest w stanie przewidzieć każdego niuansu konwencji językowych. Sztucznej inteligencji daleko jeszcze do takiej kompetencji. Należy jednak przyznać, że nie podążamy za ilością produkowanych[5] (przez każdego z nas, przez poszczególny naród) tekstów – przede wszystkim elektronicznych. Warto wziąć pod uwagę, że są one archiwizowane. Przykładowo zasób CommonCrawl[6], zawierający dane całej domeny *.ru* tylko za lata 2012 oraz 2013, liczy (hipotetycznie[7]) około 500 mld tokenów (okazów) języka rosyjskiego. SI może wspomóc pracę tłumacza. Dla nikogo już nie jest sekretem to, że SI potrafi teksty języka A oraz B urównoleglić, może zatem wspomóc pracę tłumacza nie tylko dostarczając i opracowując surowiec A, lecz także podając/podpowiadając inwariant paradygmatyczny oraz możliwe warianty referencyjne, konotacyjne, ogólnie zanurzone w sytuacyjność przebiegu wydarzeń w kontekście B. Uwzględnić należy fakt, że obecne teksty, w tym teksty Internetu, są różnorodne pod względem gatunku, lektu, stylu. Zaobserwować można wręcz lawinę tekstów – gatunków, idiolektów itd. – przeważnie w formie tekstów pisanych. Ogromna jest liczba tekstów (w tym równoległych) nieopracowanych. Warsztat związany z opracowaniem niezbadanych (nieotagowanych, nieprzeczytanych) tekstów uważamy za bardzo przydatny w pracy tłumacza.

Należy wziąć pod uwagę, że leksyka składa się z jednowyrazowców (por. niżej) i wielowyrazowców:

> *stół*, *drewniany stół*, *usiądźmy przy stol*e, *stół chirurgiczny*, *nie ma za co*, *jedziesz!*, *jedziemy z koksem!*, *ale czad!*, *wyć do księżyca*,

[4] Zob. (online) https://www.youtube.com/watch?v=DwJznO6SqyM (dostęp 12.01.2018).

[5] Świadomie nawiązujemy tu do popularnego cytatu Rolanda Barthesa: „Istnieje jednak także druga strona znaków (…): alienacja, w którą człowiek wpada, nie panując nad nadwyżką wyprodukowanych przez siebie znaków" [Barthes 1999, 14].

[6] Zob. (online) http://statmt.org/ngrams/pages/raw-data.html (dostęp 2.02.2017). W sprawie niektórych danych statystycznych tego zasobu – por. publikację [Buck i in. 2014, 3579–3584].

[7] Hipotetycznie, dlatego że niezwykle trudno jest „rozpakować" pliki tak ogromnych zasobów. Wszelkie operacje w tym operacje ekscerpcyjne, odbywają się bowiem na plikach mocno skompresowanych tychże zasobów.

a nawet formuł idiomatycznych w postaci dialogów:

– *Kto mi wózek*[8] *ukradł!?*
– *Ja! Przyznaję się bez bicia.* Itd.

Wielowyrazowce o różnej „mocy idiomatyzacji" (mocna swobodna klasa, kolokacje, idiomy – odtwarzalne jednostki leksyki), zwłaszcza w szerokiej płaszczyźnie przekładoznawstwa, wymagają obecnie współpracy, skoordynowania warsztatu badawczego leksykologa, leksykografa/frazeografa, tłumacza i informatyka. Mówiąc w bardzo dużym skrócie, przy określonej masie otagowanych równoległych frazemów (np. rosyjskich i polskich) można zautomatyzować tagowanie frazemów w translatemach właśnie w tych tekstach niezbadanych, nieotagowanych, tj. nieopracowanych. Innymi słowy, można utworzyć system, który automatycznie wyszukiwałby obiekty równoległe z tekstu równoległego, np. z tekstu specjalistycznego, ekscerpowałby i kwalifikowałby jako *frazemy potencjalne*. Rzecz jasna, ostateczna decyzja o zakwalifikowaniu takich par przekładowych do statusu reprezentatywnego należałaby do językoznawcy. Frazeografia komputerowa to dziedzina metodologii leksykograficznej, w której dokonuje się automatycznej ekscerpcji oraz analizy frazemów.

Podróż z języka A do języka B to jak podróż pomiędzy planetą Ziemią a egzoplanetą np. w Proxima Centauri. Potrzebna jest taka transgresja przekazu materiału wysłanego, by ów materiał zachować w stanie, mówiąc ogólnie, zdatnym do użycia po przybyciu „na miejsce". Podróż międzyjęzykowa to zrozumienie języka B na podstawie nawyków nabytych w języku A. Materiał wysłany do egzoplanety jest przede wszystkim przysłowiowym *punktem obserwacyjnym*. Po transgresji taki punkt (reprezentowany np. przez astronautę, urządzenia pomiarowe, SI) musiałby być zachowany, tj. musiałby nie tylko „odnaleźć się" w nowym systemie, lecz także umieć rozpoznawać nowy system na podstawie doświadczenia zdobytego jeszcze przed wyprawą w drogę – na Ziemi, w laboratorium, w sali ćwiczeń itp. Rozpoznawanie nawet najmniejszych fragmentów o klasie X w nowej rzeczywistości musi opierać się na jak największym bagażu doświadczeń (zdobytym przed podróżą, transgresją) dotyczącym identycznych lub podobnych fragmentów. Zamiarem adaptacji materiału wysłanego z Ziemi jest uniknięcie realnego błędu. Poważny błąd techniczny równa się katastrofie przedsięwzięcia. Algorytm adaptacyjny powinien uwzględniać też potencjalne błędy i rozróżniać

[8] Wersja „z wózkiem" została podsłuchana na placu zabaw: rozmawiały dwie kobiety opiekujące się dzieckiem (babcia dziecka i jej córka – matka dziecka). Naszym zdaniem dialog ten odsłania również kulisy familiolektu.

błąd od niebłędu (np. poprawnego polecenia od niepoprawnego). Dlatego zarówno podczas podróży z planety w układzie A na planetę w układzie B, jak i podróży z języka A do języka B należy „uczyć się" na *błędach potencjalnych*, by nie zaistniał realny błąd techniczny, wyrażony np. za pomocą błędnej pary przekładowej (tzw. *translatemu*) *Gaś ogień.*|*Ищи воду.* W określonej klasie sytuacji można by było (referencyjnie) uznać, że polecenie dotyczące gaszenia ognia identyczne jest z poleceniem szukania wody. Na statku mogłoby to doprowadzić nie do gaszenia ognia, tylko do szukania wody, tj. braku wyeliminowania zagrożenia. *ALARM! SYSTEM ERROR!* Podobnie byłoby w sytuacji *Ищи воду.*|*Szukaj źródła przecieku.* Tłumaczenie jest „dobre" (tj. poprawnie zaadaptowane po podróży-transgresji), kiedy uwzględnia możliwe sytuacje, do których „pasuje", tzn. racjonalnie oraz konwencjonalnie werbalizuje świat desygnatów na podstawie wiedzy dotyczącej klas sytuacji.

Paradoks adaptacji w takiej podróży polega na tym, że transgresja z języka A do języka B, tj. przenoszenie konstruktu myślowego, przewiduje sytuację, kiedy pary

Я не считаю, что это возможно.|Nie uważam, że jest to możliwe.
Я считаю, что это невозможно.|Uważam, że jest to niemożliwe.
Я не считаю, что это возможно.|Uważam, że jest to niemożliwe.
Я считаю, что это невозможно.|Nie uważam, że jest to możliwe.

mogą się nauczyć być translatemem na podstawie innych translatemów stanowiących doświadczenie pierwotne, np. *Не понимаю.*|*Nie rozumiem*., *Я не согласен.*|*Nie zgadzam się.* lub wręcz *Кот.*|*Kot*., *Нет.*|*Nie*. Dzieje się tak, dlatego że stopniowo kumuluje się zasób podobnych i odmiennych leksykalno-morfoskładniowych konstrukcji w systemie A: bagaż doświadczeń wraz z „walizką błędów": przecież translatem *Кот.*|*Kot.* pozornie w żaden sposób nie odnosi się do translatemu *Не считаю, что это возможно.*|*Nie uważam, że jest to możliwe*. Tymczasem on się odnosi. Doświadczenie „dobre" powinno iść w parze z doświadczeniem „złym". Ale jak rozpoznać, że *Nie rozumiem.* w systemie B – to analog zdania *Не понимаю.* w systemie A? Albo: czy *kot* w systemie A to również *kot* w systemie B? Przecież w systemie B (egzoplanecie przy Proxima Centauri) koty mogą nie istnieć… Przykładowo, rosyjskie zdanie *У Алисы есть кот*. musimy urównoleglić do polskiego zdania *Alicja ma kota.* W jaki sposób SI „skojarzy" te dwa zdania, nie posiadając kompetencji oraz intuicji użytkownika języka naturalnego – w określonej klasie sytuacji? Otóż okazuje się, że zdanie polskie ma zostać przetłumaczone za pomocą zdania pośredniego, ale po polsku niepoprawnego: *U Alicji jest kot* albo *Alicja, kot*. Właśnie taką (a nie inną) potencjalną wiedzę oferuje początkowo system SI oparty na dotychczasowym bagażu doświadczeń.

Automatyczny adaptator w SI generuje właśnie takie „słabe" zdania w języku B, by odpowiedni algorytm wychwycił dobre dopasowanie. Najważniejsze pytanie w tym zakresie dotyczy jakości adaptacji. Odpowiedź jest następująca: by zachować materiał „w stanie zdatnym do użycia po przybyciu na miejsce" – tak jak w przypadku podróży na Proxima Centauri – należy przede wszystkim zachować życie astronauty, urządzeń pomiarowych – punktu obserwacyjnego.

***

Niniejszy artykuł jest wynikiem badań z zakresu frazeografii komputerowej w ramach przygotowywanej monografii poświęconej zagadnieniom w danej dziedzinie. Charakter tekstu jest instruktażowy. Przedstawiamy nasz warsztat, w którym łączymy metodykę z zakresu frazematyki/frazeografii i korpusologii. W danym warsztacie osadziliśmy serie eksperymentów dotyczące urównoleglenia translatemów rosyjskich do polskich (kierunek RU→PL). Jak opracować tekst równoległy na potrzeby tłumacza, by móc tagować frazemy w obrębie par przekładowych? Najprościej skierować się do środowiska *brat*.

I. Środowisko tagowania *brat*

Środowisko narzędzia urównoleglonych zdań – to *brat v1.3*. Eksperymenty dotyczące tagowania przeprowadzane były na platformie *OS Linux Mint 18.1*. Kroki instalacji *brat v1.3* znajdują się na stronie pt. *brat installation*[9]. Kroki konfiguracji adnotacji zostały przedstawione na stronie pt. *brat annotation configuration*[10]. Tagowanie (adnotacja) to dodawanie tzw. *tagów* do wybranych obiektów (np. słów, wyrażeń, zwrotów, fraz, zdań, akapitów) w tekście cyfrowym. Przykładowo na ryc. 1 tagami są obiekty KWN_ru_Acc_FR oraz KWN_pl_Acc_FR. Nazwa relacji pomiędzy tymi dwoma tagami (TRANSGRESSION_ru_pl) także jest tagiem. Jak widzimy, środowisko *brat v1.3* umożliwia użytkownikowi posługiwanie się narzędziem do tagowania, a także udostępnia wizualnie komfortowy podgląd równoległych zdań otagowanych bądź też otagowanych obiektów (wyrazów, połączeń wyrazowych) w tych równoległych zdaniach. Architektura tagowania opiera się na grafach skierowanych[11]. Wszystkie tagi umieszczamy w osobnym pliku o nazwie *annotation.conf* (por. **Załącznik**), edytowalnym w redaktorze tekstowym. Łatwo jest więc komponować tagi w zależności od charakteru badań oraz je redagować.

[9] Zob. (online) http://brat.nlplab.org/installation.html (dostęp 23.02.2018).

[10] Zob. (online) http://brat.nlplab.org/configuration.html#configuration-basics (dostęp 23.02.2018].

[11] W sprawie terminu *graf skierowany* [Fedorushkov, Dzienisiewicz 2014, 43; Fedorushkov, Narloch 2014, 179].

Interfejs graficzny pozwala wprowadzać tagi manualnie na *otwartym* tekście, w trybie *online*. Tekst równoległy umieszczany jest bowiem na serwerze, do którego dołączane są również inne przydatne moduły.

II. Jak urównoleglić zdania? *Bleu-champ* oraz *Moses*

Dla środowiska OS Windows istnieje program „służący do urównoleglania tekstów równoległych oraz do tworzenia baz Translation Memory" – *Abbyy Aligner 1.0.6.59*[12]. Przykładowo, jeśli dobierzemy dwie wersje – polską i rosyjską – dzieła *Wojna i pokój* (np. dwa pliki *.txt), to *Abbyy Aligner* dokona urównoleglenia na podstawie określonych algorytmów i bibliotek bazowych. Eksperyment wykazał, że jakość urównoleglenia jedynie w określonej mierze jest zadowalająca. Dla X zdań rosyjskich często pojawiały się luki (brak zdań) albo zdania błędnie dopasowane.

W ramach niniejszych badań $Korpus_3$ został urównoleglony[13] na poziomie zdań za pomocą innego oprogramowania – programu *bleu-champ*[14] autorstwa Marcina Junczys-Dowmunta. Jednak dany program umożliwiający bardzo dokładne urównoleglenie wymaga dodatkowo tłumaczenia na język polski od początku. Innymi słowy, tekst dzieła *Wojna i pokój* (mimo to, że mamy wersję polską) powinien być przetłumaczony automatycznie na tekst eksperymentalny, przedurównoleglony (*proto*równoległy).

Wykorzystany został system tłumaczenia *Moses*[15]. *Moses* nie wymaga wysoce poprawnych zdań. Najważniejsze jest stworzenie doświadczenia: tego dobrego i tego złego. Wspomnijmy o Alicji i kocie. Wystarczy, by dla zdania *У Алисы есть кот* w translacie występowały *Alicja*, *kot*. albo *U Alicji kot*. albo *U Alicji jest kot*. Istota takiego podejścia polega na tym, żeby zdania nauczyć wzajemnie się kojarzyć: cechą zdania rosyjskiego i polskiego jako systemu A i B jest m.in. współwystępowanie wyrazów, tj. ważne są chociaż fragmenty takich zdań.

System *Moses* został wyuczony na podstawie 3 mln par polsko-rosyjskich zdań pobranych z korpusu napisów filmowych z roku 2016[16]. Poniżej podano przykład jakości tłumaczenia[17]:

[12] Zob. (online) https://www.abbyy.com/en-ee/aligner/ (dostęp 23.02.2018).
[13] Bazowy system operacyjny wykorzystany w niniejszej pracy to LINUX.
[14] Zob. (online) https://github.com/emjotde/bleu-champ (dostęp 23.02.2018).
[15] Zob. (online) http://www.statmt.org/moses/ (dostęp 22.02.2018).
[16] Zob. (online) http://opus.lingfil.uu.se/OpenSubtitles2016.php (dostęp 22.02.2018).
[17] Zdanie rosyjskie jest zdaniem docelowym (stąd skrót „trg" – tj. *target*, pol. *cel*). Zdanie polskie jest wyjściowe, tj. źródłowe („src", tj. *source*, pol. *źródło*).

(src)="5" > Kochanie, Dziś w nocy spełnię swoje ambicje.
(trg)="6" > Милая, сегодня моя цель будет достигнута.

(src)="6" > Odkryłem sekret życia i śmierci.
(src)="7" > W ciągu kilku godzin… powinienem stworzyć idealną ludzka istotę, jakiej świat jeszcze nie znał.
(trg)="7" > Я открыл тайну жизни и смерти, и через несколько часов я создам самого совершенного человека на свете.

Widzimy, że jakość tłumaczenia nie jest zbyt wysoka. Ale w ten właśnie sposób *bleu-champ* zdobył doświadczenie i urównoleglił polskie zdania przetłumaczone z rosyjskiego, porównując je do zdań polskich pobranych z korpusu napisów filmowych w zasobie *OpenSubtitles*.

III. Z czego składa się Korpus$_3$: literatura piękna?

Teksty rosyjskie oraz polskie tłumaczenia pobrano z różnych baz, w których rozpowszechniane są w trybie otwartego dostępu (ang. *open access*). Dobierane książki to głownie e-booki w różnych formatach: MOBI(WM), MOBI(DMR), EPUB(WM), EPUB(DMR), PDF(DRM)/PDF(ADE), FB2, ale również RTF, TXT (zob. tab.1).

Tabela 1

Zasoby Korpusu$_3$: literatura piękna

Lata powstania	Autor	Tytuł	Skrót w adnotacji
1860–1861	Iwan Turgieniew	*Ojcowie i dzieci*	@@OD
1863–1869	Lew Tołstoj	*Wojna i pokój* – I i II	@@WP
1867–1868	Fiodor Dostojewski	*Idiota*	@@ID
1937	Michaił Bułhakov	*Mistrz i Małgorzata*	@@MM
1969–1970	Wieniedikt Jerofiejew	*Moskwa – Pietuszki*	@@MP
1990–1992	Siergiej Lukjanenko	*Lord z planety Ziemia*	@@LP
1993–1994	Aleksandra Marinina	*Ukradziony sen*	@@US
1999	Wiktor Pielewin	*Generation P*	@@GP
2002–2005	Władimir Sorokin	*Lód 03 – 23000*	@@LD
2011	Ludmiła Ulicka	*Zielony namiot*	@@ZN

Źródło: opracowanie własne

Ostatecznie za pomocą programu *bleu-champ* uzyskano 78 827 par zdań (1,25 mln wyrazów rosyjskich i 1,24 mln wyrazów polskich).

IV. Efekt urównoleglenia RU→PL

W tab. 2 przedstawiamy efekt urównoleglenia zdań polskich do rosyjskich w kierunku RU→PL.

Tabela 2

Przykłady zdań urównoleglonych w podpróbce 2 (probka_2.txt w katalogu) próbki RU→PL

	TARGET	SOURCE
0.1	Не знаю, что заставило меня пойти наперекор словам Маэстро.	Nie wiem, dlaczego nie godziłem się ze słowami Maestra.
0.2	Погасил свет.	Zgasił światło.
0.3	Митенька! А Митенька! Скачи ты, Митенька, в подмосковную, – обратился он к вошедшему на его зов управляющему, – скачи ты в подмосковную и вели ты сейчас нарядить барщину Максимке-садовнику.	Mitińka, hej, Mitińka, jedźże na wieś – zwrócił się do rządcy, który wszedł na jego wezwanie – jedźże na wieś i każ natychmiast ogrodnikowi Maksymkowi, by zarządził tłokę.

Źródło: opracowanie własne

Należy zwrócić uwagę, że segmenty w tab. 2 są wymieszane. Na przykład, para przekładowa 0.1 znajduje się w utworze *Mistrz i Małgorzata*, a 0.3 należy do epopei *Wojna i pokój*. Losowość segmentów spowodowana jest procesem uczenia maszynowego: chodzi o to, by system nie był „narażony" tylko m.in. na jeden styl autorski. Takie miksowanie zdań nie jest czymś szkodliwym, tj. nie zaburza procesu badawczego. Przecież wszystkie pary przekładowe mają odpowiednią informację dotyczącą określonego źródła tekstowego w postaci tagu – por. kolumnę „Skrót w adnotacji" w tab. 1. Technicznie nietrudno zatem zdania należące do makrotagu (objaśnienia – por. niżej) @@MM zebrać w całość.

IV.1. Urównoleglenie makrotagowe: zdanie rosyjskie vs. zdanie polskie

W celu wyrażania gotowego stanu urównoleglenia używamy znaku „|||", np. *Погасил свет.|||Zgasił światło.* Wyraża on w naszym zamyśle istnienie płaszczyzn pośrednich pomiędzy płaszczyznami językowymi rozumianymi tradycyjnie jako system znaków oraz traktowanymi najczęściej syntagmatycznie: niepowtarzalne dwa makroznaki (języki jako systemy) występujące linearnie jeden po drugim, np. RU→PL, PL→RU. Płaszczyzny pośrednie to obszar metamakroznaków – tagów – występujących jako systemy makroznaków pośrednich, na zasadzie: rosyjski|RU|PL|polski, w którym RU oraz PL można interpretować jako makrotagi. Rzecz jasna, są to płaszczyzny wprowadzone sztucznie na potrzeby eksperymentalne

i w korpusie dwujęzycznym nie mają zasadniczego znaczenia. Informacje metamakrotagowe zawarte są w plikach metainformacyjnych dołączonych do każdego z tekstu podstawowego RU oraz PL.

IV.2. Niuanse urównoleglenia: zdanie rosyjskie|||zdanie polskie

Zdanie RU oraz zdanie PL to zazwyczaj jednostki minimalnego podziału w obrębie urównoleglenia. To także translatemy-zdania. Czasami może dojść do dopasowana w proporacjach: jedno zdanie vs. dwa zdania albo nawet dwa zdania vs. dwa zdania[18]. Zdarzają się także inne konfiguracje, np.

(0.1)
(trg)="55" > Он сделал вид, что на такие глупости нельзя отвечать; но действительно на этот наивный вопрос трудно было ответить что-нибудь другое, чем то, что ответил князь Андрей. – Ежели бы все воевали только по своим убеждениям, войны бы не было, – сказал он. – Это-то и было бы прекрасно, – сказал Пьер.
(src)="56" > Dał do poznania, że na takie głupstwa nie można odpowiadać, istotnie, na to naiwne pytanie trudno było odpowiedzieć inaczej, niż odpowiedział książę Andrzej. – Gdyby wszyscy wojowali tylko zgodnie z przekonaniami, toby wojen nie było – rzekł. – I to byłoby pięknie – odpowiedział Pierre [Korpus$_3$].

(0.2)
(trg)="66" > Она изображала угол тенистого сада, где поверх кустов шиповника, вырисованных с фотографической точностью, был небрежно намалеван сложный иероглиф, покрытый одинаковыми зелеными кружками. – Что это такое? – Президент на прогулке, – сказал Морковин. – Азадовский подарил для государственного настроя.
(src)="67" > Przedstawiał zakątek cienistego sadu, gdzie powyżej krzaków dzikiej róży, namalowanych z fotograficzną wiernością, niedbale nakreślono skomplikowany hieroglif, pokryty jednakowymi zielonymi krążkami. – Prezydent na przechadzce – powiedział Morkowin. – Azadowski mi to podarował dla stworzenia mocarstwowego nastroju [Korpus$_3$].

W związku z powyższym będziemy operować terminem *segment.* Natomiast w przypadku urównoleglonych segmentów, np.

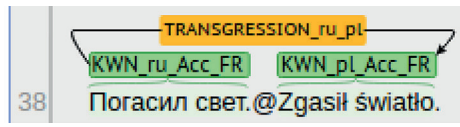
Погасил свет.|||Zgasił światło.

warto stosować określenie *segment translatemowy*, ponieważ jest umieszczony w jednym wierszu (tzw. *rekordzie*).

[18] W sprawie programów „znajdujących rozbiory składniowe wypowiedzeń", tj. tzw. *parserów* (w tym parcer o nazwie *Świgra*) [Przepiórkowski i in. 2013, 157].

W związku z tym, że mamy do czynienia z korpusem równoległym opartym na technologii urównoleglenia zdań (technicznie: serii słów od kropki do kropki), to ścieżkę urównoleglenia będziemy nazywać RU→PL: czyli od zdania RU do zdania PL. Zdanie RU oraz zdanie PL to nie tylko kolejność, tj.

(0.1) Погасил свет.|||Zgasił światło (zob. ryc. 1).



Ryc. 1. Otagowany translatem jako *segment translatemowy*: zdanie = frazem-zwrot.
Źródło: opracowanie własne

(0.2) Пьер поехал в клуб.|||Pierre pojechał do klubu. (zob. ryc. 2).

Ryc. 2. Otagowany translatem jako *segment translatemowy*: nie-frazem-zwrot w zdaniu
Źródło: opracowanie własne

W przypadku przykładu 0.1 całe zdanie – jako walencyjnie autonomiczne KWN – może stanowić wierzchołek (ang. *node*) grafu. A dla przykładu 0.2 wierzchołkami będą jedynie *поехал в клуб* oraz *pojechał do klubu* – jako osobne KWN w zdaniach RU, PL: segmenty translatemowe w obu przykładach różnią się pod względem treści tagu, a nie sposobu tagowania.

Za pomocą grafu skierowanego w kolejności RU→PL będziemy określali transgresję (definicja – por. niżej) z makroznaku A (język rosyjski, zdanie RU) do makroznaku B (język polski, zdanie PL), z których każdy ma układ zamknięty, tj. izolowany. Formalnie wyrażającym taką transgresję zdarzeniem elementarnym jest krawędź (ang. *edge*), którą określamy jako TRANSGRESSION_ru_pl.

Posługujemy się tu następującą definicją stworzoną dla potrzeb technicznych: transgresja – przekroczenie punktu obserwowanego X_1 (znajdującego się w zlokalizowanym miejscu Y_1 w izolowanym układzie systemu A) granic systemu A oraz odnalezienie się (pojawienie się) punktu obserwowanego X_1 w zlokalizowanej pozycji Y_2 w izolowanym układzie systemu B.

V. Gdzie się mieszczą tagi i kto je formułuje?

Najważniejsze pliki związane z tagowaniem to *annotation.conf*[19] oraz *logistics.ann*. Do pierwszego wprowadzane są wszystkie niezbędne tagi (znaczniki). Podczas tagowania uzupełnia się drugi plik, w którym automatycznie umieszczane są powiązania pomiędzy tagami, tj. w technicznym przełożeniu cała architektura grafowa.

Uściślijmy: plik *annotation.conf* (por. **Załącznik**) możemy redagować lokalnie; to plik zawierający tagi dla wierzchołków (wyrazów, zwrotów, wyrażeń, fraz, zdań, akapitów, tekstów, języków) oraz tagi dla krawędzi (relacje pomiędzy wierzchołkami), tzn. obejmujący dwa podstawowe zbiory tagów: wierzchołków (entities)[20] oraz krawędzi (relations). Treść tagu (wierzchołka) może być różna, np. *Czasownik*, *Rzeczownik* itd., bądź *Frazem*, *Nie-Frazem* itd. W pliku tym umieszczamy m.in. następujace tagi: KWN_ru_Acc_FR, KWN_pl_Acc_FR w translatemie *Погасил свет.|||Zgasił światło.* (por. ryc. 1).

Przykładowo: tag z treścią KWN_ru_Acc_FR oznacza, że zaznaczony obiekt w języku rosyjskim (_ru_) jest frazemem (_FR) oraz jest zwrotem werbo-nominalnym bez przyimka (KWN_), w którym czasownik występuje w związku rządu z rzeczownikiem w Bierniku (_Acc_).

Dla translatemu *Пьер поехал в клуб.|||Pierre pojechał do klubu.* (por. ryc. 2). tagami są KWN_**ru**_P_Acc_NON_FR oraz KWN_**pl**_P_Gen_NON_FR.

Przykład: tag o treści KWN_**pl**_P_Gen_NON_FR oznacza, że zaznaczony obiekt w języku polskim (_pl_) nie jest frazemem (NON_FR) oraz jest zwrotem werbo-nominalnym z przyimkiem (KWN_, _P_), w którym czasownik występuję w związku rządu z rzeczownikiem w Dopełniaczu (_Gen_).

Plik *logistics.ann* dotyczy urównoleglenia tagowego (por. ryc. 1 i 2). Początkowo ten plik jest pusty, tj. zanim nie zaczęliśmy tagować, nie ma żadnego bagażu doświadczenia, co wskazywałoby, że obiekt w zdaniu RU powiązany jest z obiektem w zdaniu PL. Po wprowadzeniu tagów do wiersza zawierającego translatem *Погасил свет.|||Zgasił światło.* (por. ryc. 1) pojawia się logistyka widoczna w tab. 3.

[19] Mówiąc o architekturze pliku *annotation.conf* (por. **Załącznik** – kolumna [relations]), należy wspomnieć, że punktem wyjścia było kilka rozwiązań technicznych opisanych w pracy: Marie-Catherine de Marneffe, Christopher D. Manning. 2008. *Stanford typed dependencies manual (Revised for the Stanford Parser v. 3.7.0 in September 2016)*. (online) http://nlp.stanford.edu/software/dependencies_manual.pdf (dostęp 20.02.2018).

[20] Por. **Załącznik**.

Tabela 3

Identyfikatory T1, T2 jako wierzchołki oraz R1 jako krawędź

ID	NAZWA TAGU	POZYCJA W PLIKU RU→PL *logistics.ann*	POKRYCIE TAGU	LOGISTYKA
T1	KWN_**ru**_Acc_FR	4220 4232	Погасил свет	
T2	KWN_**pl**_Acc_FR	4234 4248	Zgasił światło	
R1	TRANSGRESSION_ru_pl			Arg1:**T1** Arg2:**T2**

Źródło: opracowanie własne

Nazwy tagów formułowane są zatem w dowolny sposób w zależności od kierunku badań. Jednak plik logistyczny *logistics.ann* uzupełnia się automatycznie po wprowadzeniu znaczników do tekstu.

VI. Jak się dowiedzieć, czy zwrot jest frazemem?

W swoim warsztacie skupiamy się na konstrukcjach werbo-nominalnych (KWN)[21]. Próba określenia, czy zwrot jest frazemem, czy nim nie jest, stanowi osobny formalizm naukowo-badawczy. Formalizm ten polega na maksymalnie możliwej redukcji stopnia subiektywizmu podczas kwalifikowania zwrotu do klasy frazemów lub nie-frazemów, tj. okazjonalnej klasy swobodnej, generowanej na potrzeby pojedynczego opisu: np. zwroty *jeść łyżką*, *pić wiadrami* mają wysoką odtwarzalność (por. liczba wyświetleń w wyszukiwarce Google), natomiast *jeść kowszami* nie pojawia się w wyszukiwarce. A rosyjskie *есть ковшами* już tak, np. (…) *всю предыдущую неделю да и сейчас сладкое готова была* ***есть ковшами***[22].

W związku z tym stosujemy tzw. *klucz frazematyczny*[23] oraz wieloetapową weryfikację każdego zwrotu w pojedynczym kontekście, a także multikontekstowo. Często w określonej grupie kontekstów idiomatyczność oparta na modelu

[21] To właśnie KWN w systemie RU jako klasę obiektów zamierzamy nauczyć kojarzyć z klasą obiektów w systemie PL.

[22] Zob. (online): www.woman.ru/health/Pregnancy/thread/4376563/ (dostęp 5.02.2018).

[23] Opis warsztatu polegający na zastosowaniu klucza frazematycznego umieścimy w innych publikacjach. Wskażemy jedynie, że klucz frazematyczny (KF) to narzędzie kwalifikacji frazemów 3-gramowych (kolokacji, idiomów, klasy swobodnej) oraz nie-frazemów (tzw. *błędów*) przewidujące kilka etapów decyzyjnych związanych z modyfikacją albo totalną zmianą kwalifikatora. Klucz frazematyczny opiera się na *skali frazematycznej*, w której klasie swobodnej przeciwstawiane są kolokacje oraz idiomy. Te trzy klasy obiektów mają różne nasilenie idiomatyczności. Na przykład według jednej z definicji kolokacje to wyrażenia słaboidiomatyczne [Баранов, Добровольский 2014, 73]. Zob. teorie lingwistyczne dotyczące frazematyki [Chlebda 1991/2003; Федосов 2014].

przenośni może bowiem zostać udosłowniona[24]. Ostateczny klasyfikator, np. FR lub nie NON-FR, pojawia się dzięki właśnie takiej wieloetapowej weryfikacji.

Należy wspomnieć w tym miejscu, że podczas decyzji, który z tagów (FR lub NON-FR) dopasować do translatemu *Погасил свет.|||Zgasił światło*. (por. ryc. 1), musimy osobno zakwalifikować polski zwrot oraz rosyjski. Dysponujemy w tym miejscu dwoma zestawami tagów (zob. tab. 4).

Tabela 4

Zestawy tagów dla translatemu *Погасил свет.|||Zgasił światło*.

RU	PL
KWN_**ru**_Acc_FR KWN_**ru**_Acc_NON_FR	KWN_**pl**_Acc_FR KWN_**pl**_Acc_NON_FR

Źródło: opracowanie własne

Natomiast dla translatemu *Пьер **поехал в клуб**.|||Pierre **pojechał do klubu***. (por. ryc. 2) mamy do czynienia z zestawem tagów zaprezentowanym w tab. 5.

Tabela 5

Zestawy tagów dla translatemu *Пьер поехал в клуб.|||Pierre pojechał do klubu.*

RU	PL
KWN_**ru**_P_Acc_FR KWN_**ru**_P_Acc_NON_FR	KWN_**pl**_P_Gen_FR KWN_**pl**_P_Gen_NON_FR

Źródło: opracowanie własne

Trzeba zdawać sobie sprawę z tego, że obiekt w RU, będący FR, nie musi być FR w PL i odwrotnie – tak jak przypadku *jeść kowszami* vs. *есть кошвами*. Polski przykład oznaczylibyśmy klasyfikatorem NON_FR, a rosyjski – FR.

Jeden z etapów weryfikacji obejmowałby czynność weryfikacyjną, polegającą na sprawdzeniu faktu rejestracji leksykograficznej, np. dla *Погасил свет.* odnaleziono przykład w źródle leksykograficznym: por. ryc. 3, w której zaprezentowano artykuł hasłowy dla hasła *погасить*.

[24] Por. połączenie *czarna owca* w zdaniach *W naszym biurze zaczęła pracować jakaś nowa czarna owca*. oraz *Na łące pasły się biała i czarna owca*. W drugim zdaniu kongruencja członów połączenia *czarna owca* nie jest obarczona modelem metafory. Podobnie z wyrażeniem *jeść z ręki*$_1$: w określonej grupie kontekstów werbalizuje sytuację dosłownie: że ‘ktoś komuś je coś fizycznie z ręki’ – np. *Wiewiórka uczy się jeść z ręki*. W innej grupie kontekstów *jeść z ręki*$_2$ zawiera model przenośni (oznacza ‘być podporządkowanym’) i jest frazemem. Nie oznacza to jednak, że *jeść z ręki*$_1$ frazemem nie jest. Mają bowiem one różne statusy frazematyczne. Zwrot *jeść z ręki*$_1$ odnosimy do idiomatycznie słabszych – kolokacji, a *jeść z ręki*$_2$ – do idiomatycznie mocniejszych – idiomów. Por. traktowanie terminu *kolokacja* [Баранов, Добровольский 2014, 73] w przyp. 25.

погасить
погасить долги • существование / создание, прерывание, решение, компенсация
погасить свет • действие, прерывание
свет погасить • действие, прерывание

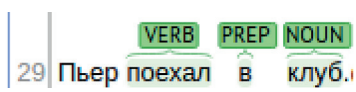
Ryc. 3. Artykuł *погасить*[25]
Źródło: (online) http://abstrnoun.academic.ru/3588/погасить (dostęp 5.02.2018)

Dzięki dodatkowym czynnościom weryfikacyjnym (występowania zwrotu RU oraz PL w różnych kontekstach, modyfikacji znaczenia zwrotu, współwystępowania wariantów) została podjęta decyzja, że dany zwrot jest frazemem zarówno w RU, jak i PL.

VII. Uwypuklenie wymiarów: gramatyczny vs. frazematyczny

Przyjęliśmy, że KWN w translatemie *поехал в клуб*|||*pojechał do klubu* należą do klasy swobodnej (NON_FR). Niefrazematyczna konstrukcja została przetłumaczona również niefrazematycznie (NON_FR). Różnica polega jedynie na realizacji aktantu werbalizatora: w języku rosyjskim wraz z przyimkiem *в* figuruje Accusativus, w polskim – przyimek *do* w Genetivus. Ostatecznie wyraża się to w grafie zawierającym wierzchołkowo-krawędziową zbitkę tagową (por. ryc. 2).

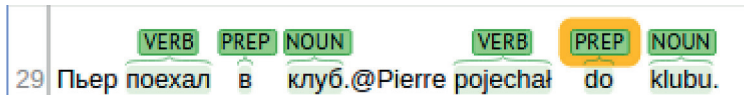
Uwypuklenie wymiaru gramatycznego ma jedynie charakter „lokalny", tj. wprowadzenie kategorii Część mowy (POS) miałoby sens w granicach – generalnie rzecz biorąc – tylko jednego języka. Jeśli w przypadku przykładu KWN *поехал в клуб* w obszarze RU nie mielibyśmy nic przeciwko tagowaniu p e ł n e m u (zob. ryc. 4),



Ryc. 4. POS dla *поехал в клуб*
Źródło: opracowanie własne

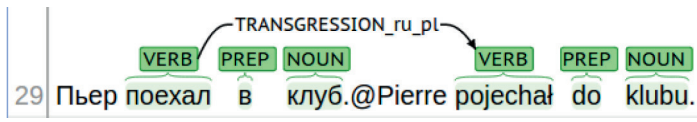
to w sytuacji z translatem (PL) nie miałoby to najmniejszego sensu ze względu na holistyczność tagową (zob. ryc. 5).

[25] Zob. [Бирюк et al. 2008].

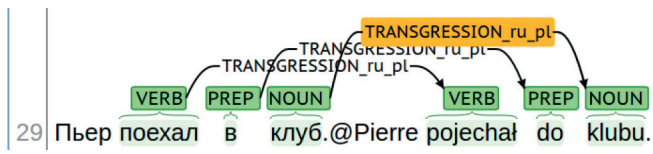


Ryc. 5. POS dla *поехал в клуб|||pojechał do klubu*
Źródło: opracowanie własne

Wymagałoby to wprowadzenia rozróżnienia tagów POS dla RU i PL, np. VERB_RU, VERB_PL, NOUN_RU, NOUN_PL itd. Ponadto nadnarzucenie dodatkowych tagów relacyjnych (np. TRANSGRESSION_ru_pl) wiązałoby się z regulacją błędną w stosunku do krawędzi typu V (RU) → V (PL), Prep (RU) → Prep (PL), Noun (RU) → Noun (PL), w których figurowałby ten sam tag (por. ryciny 5.1 i 5.2).



Ryc. 5.1. POS dla *поехал в клуб|||pojechał do klubu*
Źródło: opracowanie własne



Ryc. 5.2. POS dla *поехал в клуб|||pojechał do klubu*
Źródło: opracowanie własne

Innym problemem jest „miksowanie" tagów dla jednowyrazowców (pojedynczych wyrazów z przypisanym pojedynczym tagiem) i wielowyrazowców w tym samym zdaniu. W praktyce przekłada się to na błąd logistyczny, przy którym wielowyrazowcom nadawany jest tag związany z charakterystyką morfoskładniową i frazematyczną (KWN_ru_P_Acc_NON_FR), a jednowyrazowcom tag dotyczący jakieś kategorii gramemowej (np. rodzaj żeński, liczba pojedyncza) albo gramatycznej (POS) – zob. ryc. 6.

Ryc. 6. Błąd tagowania: tag dla pojedynczego wyrazu (jednowyrazowca) vs. tag dla kilku wyrazów (trzywyrazowca)
Źródło: opracowanie własne

Wskazujemy zatem na zasadę oszczędzania wytycznych do analizy adaptacyjnej – to właśnie niepodawanie wszystkich drobnych danych w języku A oraz B jest swoistym rozwiązaniem technicznym obejmującym zarówno umiarkowaną indukcję, jak syntetyzm. Lepiej utworzyć bowiem dwa takie same korpusy równoległe i różnie je otagować niż tagować wszystko, co się da lokalnie. Zasada ta sprawdza się w przypadku braku ekwiwalentów konstrukcyjnie (tj. morfoskładniowo) zbliżonych: np. w języku A figuruje jakiś zwrot, a w języku B brakuje ekwiwalentu i tłumaczenie odbywa się w sposób analityczny lub składniowo odmienny, por.

(…) скачи ты в подмосковную и **вели** ты сейчас **нарядить барщину** Максимке-садовнику.||| (…) jedźże na wieś i **każ** natychmiast ogrodnikowi Maksymkowi, **by zarządził tłokę**.

Podsumowanie

Odwróćmy kolejność czynności naszkicowanych w niniejszym tekście:

1. Ustalenie: dobór wymiaru tagowania (np. tagowanie wielowyrazowców w aspekcie frazematycznym).
2. Ustalenie: dobór klucza frazematycznego oraz jego zastosowanie do wielowyrazowców we wcześniej przeanalizowanym tekście równoległym.
3. Ustalenie: utworzenie tagów (znacznków) dla obiektów (frazemów) oraz znaczników dla relacji w formalizmie urównoleglenia.
4. Ustalenie: dobór formalizmu urównoleglenia, np. *Zdanie RU.|||Zdanie PL.* albo *Fragment tekstu RU.|||Treściowo identyczny fragment PL*.
5. Ustalenie: dopasowanie tagów w obszarze makrotagów, np. RU, PL, lekt (np. literatura, idiolekt), przynależność do tekstu A, B, C itd., przynależność do stylu A, B, C itd.
6. Ustalenie: dobór próbek. Czy wybrać zdania równoległe jednego tekstu w próbce czy „wymieszać" zdania równoległe z kilku tekstów równoległych?
7. Znalezienie tekstów równoległych.
8. Wyszkolenie systemu urównoleglenia.
9. Urównoleglenie.
10. Praktyczne zastosowanie etapów 1–6.

Życzymy udanej podróży!|||Желаем счастливого пути!

Bibliografia

Baranov Anatolij Nikolaevič, Dobrovol′skij Dmitrij Olegovič. 2014. *Osnovy frazeologii (kratkij kurs). Učebnoe posobie*. Moskva: Flinta [Баранов Анатолий Николаевич, Добровольский Дмитрий Олегович. 2014. *Основы фразеологии (краткий курс). Учебное пособие.* Москва: Флинта].

Birûk Ol′ga Leonidovna, Gusev Valentin Ûr′evič, Kalinina Elena Ûr′evna. 2008. *Slovar′ glagol′noj sočetaemosti nepredmetnyh imën*. Institut russkogo âzyka im. V.V. Vinogradova RAN, Nacional′nyj korpus russkogo âzyka. (online) http://abstrnoun.academic.ru (dostup 5.02.2018) [Бирюк Ольга Леонидовна, Гусев Валентин Юрьевич, Калинина Елена Юрьевна. 2008. *Словарь глагольной сочетаемости непредметных имён.* Институт русского языка им. В.В. Виноградова РАН, Национальный корпус русского языка. (online) http://abstrnoun.academic.ru (доступ 5.02.2018).

Barthes Roland. 1999. *Imperium znaków*. Warszawa: Wydawnictwo Kr.

Buck Christian, Heafield Kenneth, Van Ooye Bas. 2014. *N-gram Counts and Language Models from the Common Crawl*. W: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Red. Calzolari N. (Conference Chair), Choukri Kh., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. Reykjavik: European Language Resources Association (ELRA): 3579–3584. (online) http://www.lrec-conf.org/proceedings/lrec2014/pdf/1097Paper.pdf (access 12.01.2017).

Chlebda Wojciech. 2003. *Elementy frazematyki: wprowadzenie do frazeologii nadawcy*. Wyd. 2. Łask: Oficyna Wydawnicza Leksem.

Fedorushkov Yury, Dzienisiewicz Daniel. 2014. *Automatyzacja wizualizacji grafowej synonimów dla potrzeb dydaktyki języków obcych (na przykładzie przymiotników rosyjskich z prefiksem без-/бес-)*. „Kultury Wschodniosłowiańskie – Oblicza i Dialog. Białoruś. Rosja. Ukraina" nr 4: 35–48.

Fedorushkov Yury, Narloch Andrzej. 2014. *Prolegomena do dydaktycznej prezentacji konceptu językowego w wizualizacji grafowej (na przykładzie rosyjskiego konceptu БЕЛЫЙ)*. „Studia Rossica Gedanensia" nr 1: 179–208.

Fedosov Oleg Ivanovič. 2014. *Perevod kvazifrazem (k probleme izučeniâ slaboidiomatičnyh ustojčivyh sočetanij)*. W: *Frazeologia a przekład. Materiały z konferencji językoznawczej (4–6 września 2011 r., Opole)*. Red. Chlebda W. Opole: Wydawnictwo Uniwersytetu Opolskiego: 359–366 [Федосов Олег Иванович. 2014. *Перевод квазифразем (к проблеме изучения слабоидиоматичных устойчивых сочетаний).* W. *Frazeologia a przekład. Materiały z konferencji językoznawczej (4–6 września 2011 r., Opole)*. Red. Chlebda W. Opole: Wydawnictwo Uniwersytetu Opolskiego: 359–366.]

de Marneffe Marie-Catherine, Manning Christopher D. 2008. *Stanford typed dependencies manual (Revised for the Stanford Parser v. 3.7.0 in September 2016)*. (online) http://nlp.stanford.edu/software/dependencies_manual.pdf (dostęp 20.02.2018).

Przepiórkowski Adam, Skwarski Filip, Hajnicz Elżbieta, Patejuk Agnieszka, Świdziński Marek, Woliński Marcin. 2013. *Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego*. „Polonica" nr XXXIII: 157–175.

Raževa Elena Ivanovna. 2006. *Limerik: neperevodimaâ igra slov ili perevodimaâ igra formy?* W: *Logičeskij analiz âzyka. Konceptual'nye polâ igry*. Red. Arutûnova N.D. Moskva: Indrik: 327–335. (online) http://ec-dejavu.ru/l/Limerick.html (dostup 11.12.2017) [Ражева Елена Ивановна. 2006. *Лимерик: непереводимая игра слов или переводимая игра формы?* W: *Логический анализ языка. Концептуальные поля игры*. Red. Арутюнова Н.Д. Москва: Индрик: 327–335. (online) http://ec-dejavu.ru/l/Limerick.html (доступ 11.12.2017)].

Summary

Prolegomena for tagging of phrasemes in a parallel Russian-Polish corpus (literature) in translation studies

This article considers tagging methods for parallel Russian-Polish phrasemathic objects. In particular, an opinion about the annotation tool *brat v1.3.*is given. This online tool offers a palette of possibilities for classifying words and phrases in parallel texts. Working with this software is largely simplified by a user-friendly interface, and therefore working with the corpus does not cause difficulties for philologists and translators who do not have programming skills. As an example of such a classification, the layout of the metadata system for tagging Russian and Polish parallel phrasemes is described. These resources allow experience to be gathered and concurrent objects to be categorized in the workshop of a translator. As an example, the article presents the tagging of Verb-Noun of the text classified as collocation phrasemes, for example, *погасить свет*. The status of Verb-Noun constructions is also discussed, which, according to a number of factors, relate to autonomous phrases, although with the status of "free compatibility", for example, *поехать в клуб*. A number of recommendations is proposed for the configuration of parallel texts at the level of single sentences.

Key words: annotation tool *brat v1.3.*, tags for phrasemes, Verb-Noun constructions, parallelization of Russian and Polish sentences, parallel corpora

Kontakt z Autorem:
jerfed@amu.edu.pl

Załącznik 1. *annotation.conf*

```
[entities]
KWN_ru
	KWN_ru_Gen_FR
	KWN_ru_Gen_NON_FR
	KWN_ru_Dat_FR
	KWN_ru_Dat_NON_FR
	KWN_ru_Acc_FR
	KWN_ru_Acc_NON_FR
	KWN_ru_Ins_FR
	KWN_ru_Ins_NON_FR
	KWN_ru_Pre_FR
	KWN_ru_Pre_NON_FR
KWN_ru_P
	KWN_ru_P_Gen_FR
	KWN_ru_P_Gen_NON_FR
	KWN_ru_P_Dat_FR
	KWN_ru_P_Dat_NON_FR
	KWN_ru_P_Acc_FR
	KWN_ru_P_Acc_NON_FR
	KWN_ru_P_Ins_FR
	KWN_ru_P_Ins_NON_FR
```

```
[relations][26]
<TOKEN>=<ENTITY>

TRANSGRESSION_ru_pl  Arg1:<TOKEN>, Arg2:<TOKEN>
TRANSGRESSION_pl_ru  Arg1:<TOKEN>, Arg2:<TOKEN>
root   Arg1:<TOKEN>, Arg2:<TOKEN>
	dep    Arg1:<TOKEN>, Arg2:<TOKEN>
	aux    Arg1:<TOKEN>, Arg2:<TOKEN>
	arg    Arg1:<TOKEN>, Arg2:<TOKEN>
		comp    Arg1:<TOKEN>, Arg2:<TOKEN>
			obj    Arg1:<TOKEN>,
Arg2:<TOKEN>
		subj    Arg1:<TOKEN>, Arg2:<TOKEN>
	cc     Arg1:<TOKEN>, Arg2:<TOKEN>
	conj   Arg1:<TOKEN>, Arg2:<TOKEN>
	expl   Arg1:<TOKEN>, Arg2:<TOKEN>
	mod    Arg1:<TOKEN>, Arg2:<TOKEN>
		amod    Arg1:<TOKEN>, Arg2:<TOKEN>
		det     Arg1:<TOKEN>, Arg2:<TOKEN>
		nn      Arg1:<TOKEN>, Arg2:<TOKEN>
		num     Arg1:<TOKEN>, Arg2:<TOKEN>
```

[26] Zob. [de Marneffe, Manning 2008].

KWN_ru_P_Pre_FR prep Arg1:<TOKEN>, Arg2:<TOKEN>
KWN_ru_P_Pre_NON_FR punct Arg1:<TOKEN>, Arg2:<TOKEN>
KWN_pl
KWN_pl_Gen_FR
KWN_pl_Gen_NON_FR
KWN_pl_Dat_FR
KWN_pl_Dat_NON_FR
KWN_pl_Acc_FR
KWN_pl_Acc_NON_FR
KWN_pl_Ins_FR
KWN_pl_Ins_NON_FR
KWN_pl_Pre_FR
KWN_pl_Pre_NON_FR
KWN_pl_P
KWN_pl_P_Gen_FR
KWN_pl_P_Gen_NON_FR
KWN_pl_P_Dat_FR
KWN_pl_P_Dat_NON_FR
KWN_pl_P_Acc_FR
KWN_pl_P_Acc_NON_FR
KWN_pl_P_Ins_FR
KWN_pl_P_Ins_NON_FR
KWN_pl_P_Pre_FR
KWN_pl_P_Pre_NON_FR
NON_KWN_FR
NON_KWN_NON_FR
PodOrz_ru_FR
PodOrz_ru_NON_FR
KUM_FRAZ_ru
KUM_FRAZ_pl
PodOrz_pl_FR
PodOrz_pl_NON_FR
ZERO_WORD
ADJ
ADP
ADV
CONJ
DET
NOUN
NUM
PRON
PRT
VERB
X
PUNCT
PREP
ZERO