Danuta Roszko
Polska Akademia Nauk

Roman Roszko
Polska Akademia Nauk

Polsko-litewskie korpusy równoległe. Elementy anotacji semantycznej z zakresu modalności możliwościowej i kwantyfikacji zakresowej

Polish-Lithuanian parallel corpora: Elements of the semantic annotation related to hypothetical and imperceptive modalities and scope quantification.

Abstract

The authors present two Polish-Lithuanian parallel corpora: (1) experimental *EKorpPL-LT* and (2) *KorpPL-LT_CLARIN*. *EKorpPL-LT* is the first extended bilingual Polish-Lithuanian corpus whose resources have been divided into two subcorpora: parallel and comparable. The parallel subcorpus is widely applied in contrastive studies carried out at the Institute of Slavic Studies, Polish Academy of Sciences by the Corpus Linguistics and Semantics Team. *Parallel EKorpPL-LT* contains various texts being mutual translations between these two languages. *KorpPL-LT_CLARIN* is based on vast fragments of translations of fiction writings and specialist texts. It is created within the framework of the Polish scientific consortium being a section of the pan-European research infrastructure called CLARIN. For both corpora, basic applications established by their authors are presented. Next, the authors portray the archaic nature of the Lithuanian language, which is of benefit to the structure of multilingual corpora. For this purpose the basic assumptions of semantic categories such as (a) definiteness/indefiniteness, (b) modality (b1) hypothetical and (b2) imperceptive are described. Next, under the distinguished categories and on the basis of the Lithuanian language distinctive features, the possibility to extend the description of the Polish corpora resources is discussed.

The authors present some examples of a new semantic annotation (developed by Violetta Koseska and Roman Roszko – for scope quantification and Danuta Roszko and Roman Roszko – for modality). The authors distinguish the following three semantic units:

- A neutral degree (I1) and an enhanced degree (I2) of imperceptiveness,
- A degree of the lowest probability (H1), particular degrees of growing probability (H2–H5) and a degree of the highest probability (H6) of hypothetical modality,
- Uniqueness, existentiality (E1), real existentiality, habitual universality and real universality (categories of scope quantification).

The authors assume that the conservative nature of the Lithuanian language, manifesting itself in (i) the stability of forms, (ii) relations between the form and its function, (iii) narrowed specialization of forms, much more advanced than in the Polish language, not only allows to extend the description of the resources, but also considerably affects the development of linguistics and all applied sciences based on language (such as the process of teaching the language, traditional and machine translation etc.).

Słowa kluczowe: korpus równoległy, język polski, język litewski, anotacja semantyczna, automatyzacja przekładu
Keywords: parallel corpus, Polish, Lithuanian, semantic annotation, automation of translation / machine translation.

1. Dwa polsko-litewskie korpusy równoległe

Polsko-litewskie zasoby tekstowe są częścią wielu korpusów wielojęzycznych (np. Emea, InterCorp, KDE4, Opus, ParaSol i in.). Jednak w tym artykule zamierzamy zwrócić uwagę na dwa korpusy, które zaplanowano z myślą o konkretnych zastosowaniach. Pierwszy z nich to eksperymentalny korpus polsko-litewski (dalej EKorpPL-LT). Drugi z nich to polsko-litewski korpus równoległy (dalej KorpPL-LT_CLARIN), powstający w ramach zadań polskiego konsorcjum naukowego, będącego częścią ogólnoeuropejskiej infrastruktury badawczej CLARIN.

1.1 Eksperymentalny korpus polsko-litewski EKorpPL-LT

Korpus EKorpPL-LT był intensywnie rozwijany w latach 2010–2012. Inicjatorami i głównymi wykonawcami tego korpusu są autorzy tego artykułu. Kiedy jednak ci sami autorzy przystąpili do prac nad nowym korpusem KorpPL-LT_CLARIN, rozwój EKorpPL-LT został wstrzymany. Jednak w krótkim czasie okazało się, że rozwijanie korpusu eksperymentalnego jest uzasadnione i celowe. Zadecydowały o tym dwa – jak się później okazało – wzajemnie wykluczające się czynniki. Pierwszy z nich był związany z przeświadczeniem, że nowy korpus KorpPL-LT_CLARIN – jako „spadkobierca" korpusu eksperymentalnego – będzie doskonalszym narzędziem w ręku językoznawcy i tłumacza. Drugi zaś

czynnik, to przyjęte w korpusie KorpPL-LT_CLARIN założenie włączenia do zasobów korpusowych tylko tekstów licencjonowanych, które spowodowało istotne ograniczenia liczby potencjalnych do zamieszczenia w korpusie tekstów. Uzyskanie licencji na włączenie tekstu (nawet jego fragmentu) do korpusu jest zadaniem trudnym do spełnienia. Dlatego w roku 2014 wznowiono prace nad EKorpPL-LT. Obecne parametry korpusu to: 2 006 021 słowoform tekstów beletrystycznych i 14 210 323 słowoform współczesnych tekstów specjalistycznych, będących w dużej mierze tłumaczeniami wzajemnymi (tj. z języka polskiego na litewski lub odwrotnie). Szczególną uwagę w doborze materiału skupiono na tekstach specjalistycznych. Zadbano, by w EKorpPL-LT znalazły się reprezentatywne dla poszczególnych dziedzin teksty, charakteryzujące się poprawną stylistyką i terminologią w obu wersjach językowych (por. tabela 1.) z przykładami tekstów, które nie spełniły wymagań autorów EKorpPL-LT.

Przykład tekstu w języku		**Przyczyna odrzucenia tekstu**
polskim	**litewskim**	
Zawartość frazy stałej NMR 20 st. C 20–28%	Kietos frazės NMR skaičius 20°C 20–28%	Obcy skrótowiec NMR (por. ang. nuclear magnetic resonance) w obu językach dot. spektroskopii magnetycznego rezonansu jądrowego. Zapis polski *20 st. C* zamiast *20°C*.
Wymagane jest oprzyrządowanie laboratoryjne do analizy chemicznej substancji badanej i produktów przemian (np.: chromatografia gazowo-cieczowa (GLC), spektroskopia masowa (MS), magnetyczny rezonans jądrowy (NMR) itp.), włączając w to układy do wykrywania substancji chemicznych znakowanych i nieznakowanych radioizotopami.	Reikia turėti laboratorinius bandomosios medžiagos ir transformavimo produktų analizės prietaisus (pvz., dujų ir skysčių chromatografijos (DSC), masių spektroskopijos (MS), magnetinio branduolių rezonanso (MBR) ir t. t. įrangą), atitinkamas aptikimo sistemas žymėtosioms arba nežymėtosioms medžiagoms analizuoti.	Obce skrótowce w tekście polskim (*GLC*, *MS* i *NMR*). Brak gramatycznego i kontekstowego powiązania w tekście polskim fragmentu w nawiasach *(np.: chromatografia gazowo-cieczowa (GLC), spektroskopia masowa (MS), magnetyczny rezonans jądrowy (NMR) itp.)* z poprzedzającym go kontekstem. Brak formy *urządzenia* w tekście polskim w przytoczonym wyżej fragmencie w nawiasach.

Tabela 1. Wyjątki z tekstów, które nie zostały włączone do zasobów korpusowych EKorpPL-LT ze względów formalnych

Dziedziny najlepiej reprezentowane w korpusie EKorpPL-LT to: przemysł energetyczny, stoczniowy, budowlany, spożywczy, chemiczny, farmaceutyczny, naftowy, biotechnologiczny, metalurgiczny, bankowość, rachunkowość, bezpieczeństwo i higiena pracy, ochrona środowiska, medycyna, prawo i sądownictwo, ustawodawstwo unijne, urządzenia elektrotechniczne (użytku domowego

i przemysłowe), komunikacja w ruchu lądowym i powietrznym, wykaz działalności i towarów. Jak można zauważyć, uwaga twórców EKorpPL-LT skupia się na tekstach zawierających najnowszą terminologię, która nie została uwzględniona w żadnym polsko-litewskim słowniku drukowanym czy elektronicznym. W związku z tym oczywisty staje się cel powstania tego korpusu, mianowicie budowa pamięci tłumaczeniowych[1] (por. tabela 2.) oraz jego wykorzystanie w pracach nad planowanym w Instytucie Slawistyki PAN wielojęzycznym słownikiem nowej generacji.

Litera	Termin polski	Termin litewski
B	badania dodatkowe	kiti bandymai
	badania makroskopowe	makroskopinis tyrimas
	badania radiograficzne	radiografinis bandymas
	badania wizualne	apžiūrimoji kontrolė
	blacha	plokštė
P	próba łamania	laužimo bandymas
	próba rozciągania	tempimo bandymas
	próba zginania	lenkimo bandymas

Tabela 2. Wyciąg z pamięci tłumaczeniowej. Terminy z zakresu spawalnictwa (fragmenty)

1.2 Polsko-litewski korpus równoległy KorpPL-LT_CLARIN[2]

Ten korpus, o zaplanowanej na rok 2016 objętości przekraczającej sześć milionów słowoform, bazuje na obszernych fragmentach licencjonowanych utworów beletrystycznych, różnorodnych tekstach specjalistycznych, w tym pochodzących z domeny publicznej. KorpPL-LT_CLARIN wpisuje się w standardy obowiązujące w lingwistyce korpusowej. Zaplanowano ręczne naniesienie eksperymentalnej anotacji semantycznej (dotyczącej kwantyfikacji zakresowej na poziomie fraz nominalnej i werbalnej) w tekstach ciągłych do łącznej objętości 4 000 zdań. Szerzej na temat stosowanej w KorpPL-LT_CLARIN anotacji semantycznej (por. Koseska-Toszewa, 2013; Roszko, D., Roszko, R., 2013;

1 Autorzy zawodowo zajmują się tłumaczeniami. W pracy translatorskiej kierują się zasadą konsekwentnego stosowania adekwatnej i spójnej terminologii oraz unikania ponownego tłumaczenia identycznych lub podobnych fragmentów. W tym celu tworzą pamięci tłumaczeniowe oraz stosują oprogramowanie wspomagające tłumaczenie.

2 KorpPL-LT_CLARIN powstaje w ramach zadań polskiego konsorcjum naukowego, będącego częścią ogólnoeuropejskiej infrastruktury badawczej CLARIN (nr projektu 131241). Liderem konsorcjum jest Politechnika Wrocławska (kierownik – Maciej Piasecki). KorpPL-LT_CLARIN jest zadaniem włączonym do modułu 9, realizowanym przez Zespół Lingwistyki Korpusowej i Semantyki Instytutu Slawistyki PAN (kierownik Zespołu i koordynator ze strony IS PAN – Violetta Koseska-Toszewa, główni wykonawcy: Danuta Roszko i Roman Roszko).

Koseska, Roszko, 2015). Docelowym odbiorcą tego korpusu są przedstawiciele szeroko pojętych nauk humanistycznych.

2. Zachowawczy charakter języka litewskiego i wynikające z tego faktu korzyści, mogące mieć zastosowanie w rozbudowie samych korpusów oraz w doskonaleniu algorytmów przekładu maszynowego

Jak dotąd, oczywiste dla bałtystów i indoeuropeistów cechy języka litewskiego nie zostały dostrzeżone jako potencjalne źródło sprzyjające (a) automatyzacji prac nad budową wielojęzycznych korpusów oraz (b) identyfikacji znaczeń doskonale odzwierciedlonych na płaszczyźnie formalnej. Aby umożliwić zrozumienie zachowawczego charakteru języka litewskiego, omówmy prostą polsko-litewską parę odpowiedników: *syn – sūnus* 'syn'. Zachowawczy charakter języka litewskiego pozwala w sposób prosty wyprowadzić nawet ze współczesnej litewskiej formy *sūnus* postać prasłowiańską i późniejszą polską *syn*. W slawistyce oczywiste jest przewartościowanie długiego ie. ū w prasłowiańskie *y*, które bez zmian jest obserwowane we współczesnej polskiej formie *syn*. Następnie, powodowana prawem sylab otwartych (inaczej prawem wzrastającej dźwięczności) fleksja mianownikowa *-s* została zredukowana (utracona), zaś wygłosowe krótkie *u* uległo zmianie w jer (ъ), który – będąc w absolutnym wygłosie – również uległ zanikowi. Cały proces można przedstawić w ciągu: **sūnus* → **sūnu* → *synъ* → *syn*. Dla formy dopełniaczowej można przedstawić następujący proces: **sūnaus* → **sūnau* → *synu* → *syna*, gdzie ponadto stwierdzamy dobrze w literaturze opisany proces monoftongizacji dyftongu **au* do *u*, a następnie zmianę fleksji dopełniaczowej *u* (pod wpływem odmiany na *a* krótkie) do postaci *a*.

Tym razem nieco rozbudujmy wyjściową parę do postaci *słodki syn – saldus sūnus* 'słodki syn'. W tym wypadku dochodzą kolejne zmiany, które doprowadziły nie tylko do utraty pierwotnej postaci w wyniku zmian fonetycznych, lecz również utajniły inne procesy, takie jak dodanie elementu *-k-*[3] (por. również bułg. сладък, ros. сладкий), czy rozbudowę o kolejny element tym razem pochodzenia zaimkowego **-jis*, który całkowicie zlany z rdzeniem przyczynił się do ukształtowania zupełnie nowej odmiany przymiotników w języku polskim. Dawne znaczenie wnoszone przez kontynuant **-jis* zostało zupełnie zatarte we współczesnej polszczyźnie[4]. Przedstawiona zatem wyżej para ekwiwalentów *słodki syn – saldus sūnus* powinna (tylko z formalnego punktu widzenia) przybrać następującą postać: *słodki syn – saldusis sūnus*, gdzie w formie

3 Przypuszcza się, że wzbogacanie form o element *-k-* w językach słowiańskich było zabiegiem czysto formalnym, mającym rozróżnić przymiotniki i rzeczowniki, por. chociażby współczesną postać niem. *süß* bez *-k-*, a także inną analogiczną polsko-litewską odpowiedniość: *gorzki – kartus*.

4 Por. archaiczne polskie postacie *zdrów* (a *zdrowy*), *rad* (a *rady*) i in., w których nie stwierdza się kontynuantu dawnego **-jis*.

litewskiej wyraźnie obserwujemy element *-jis*, por. *saldus+jis* → *saldusis*. Na płaszczyźnie znaczeniowej litewskie *saldusis* w opozycji do *saldus* jest wykładnikiem znaczeń kwantyfikacyjnych jednostkowości ('ten słodki …') i ogólności zwyczajowej 'zazwyczaj każdy słodki …', por. analogiczne zjawisko w językach rodzajnikowych oraz zaobserwowane w nich funkcje tzw. rodzajnika określonego (Karolak, 2001).

Przedstawione w wielkim skrócie polsko-litewskie odpowiedniości dokumentują znaczne zmiany form polskich leksemów. Jak można zauważyć chociażby w formie *syn*, ulega zanikowi dawna fleksja (np. mianownikowa *-s*), także tematyczne *-u*. Natomiast w formie *słodki* odnotowujemy istotne w naszych rozważaniach zlanie się form (zatarcie granic między morfemami). Dawne **-jis* zostało wkomponowane w postać poprzedzającej formy. Pierwotne znaczenie wnoszone przez **-jis* zostało utracone i faktycznie stało się wyznacznikiem odmiany przymiotnikowej.

Zatem uproszczenia i zmiany w strukturze formalnej polskich leksemów (w tym przede wszystkim rozmycie granic między morfemami), doprowadzające do zachwiania stabilności wnoszonych przez poszczególne morfemy znaczeń, prowadzą do utraty bezpośredniego powiązania morfemu/formantu ze znaczeniem oraz umacniania się struktur nieprzejrzystych formalnie. Zacieranie granic między morfemami (również gramatycznymi) narusza prostą odpowiedniość formy i jej znaczenia – tym samym ten proces prowadzi do zahamowania, zaniku pewnych znaczeń oraz wykształcenia nowych, jak można zaobserwować, nie zawsze konsekwentnego w całym paradygmacie, por. chociażby polską kategorię deprecjacji ograniczoną do wybranych leksemów rzeczownikowych oraz form przypadkowych.

Formy litewskie w odróżnieniu od polskich są stabilne. Wyniki analiz kategorii semantycznych w języku litewskim, (por. Roszko, R., 1993, 2004; Roszko, D., 2006, 2015) ujawniają również inną cechę języka litewskiego – wraz z zachowaniem pierwotnej struktury formalnej wyrazu zostaje zachowana łączność (w tym stabilność) między poszczególnymi formantami a ich znaczeniem. Tę właśnie cechę współczesnej litewszczyzny zamierzamy wykorzystać w identyfikacji nieujawnionych na płaszczyźnie formalnej polszczyzny znaczeń i włączenie takim sposobem ustalonych znaczeń do opisu, definiowanego tu jako anotacja semantyczna.

3. Przykłady anotowanych struktur semantycznych

3. 1 Kwantyfikacja zakresowa – semantyczna kategoria określoności-nieokreśloności

Jest to kategoria zdaniowa (tj. dotyczy zarówno frazy nominalnej jak i werbalnej) z wyróżnioną opozycją jednostkowości: niejednostkowości. Określoność odpowiada treściom jednostkowości (z podziałem na jednostkowość elementu i zbioru), natomiast nieokreśloność – niejednostkowości, obejmującej znaczenia egzystencjalności (ograniczonej i właściwej) i ogólności (zwyczajowej/ograniczonej i właściwej), (por. Koseska-Toszewa, 1982; Косеска-Тошева, Гаргов, 1990; Roszko, R., 2004; Roszko, D., 2015). W definicji trzech podstawowych pojęć wykorzystano powszechnie znane znaczenia kwantyfikatorów logicznych (kwantyfikatora szczegółowego i ogólnego) oraz jota-operatora. W opisie wykładników tej kategorii posłużono się również pojęciem niedopowiedzenia kwantyfikacyjnego, zauważonym przez Ajdukiewicza (1965). Szczegóły dotyczące semantycznej kategorii określoności-nieokreśloności oraz opis jej wykładników (leksykalnych, morfologicznych i składniowych) zarówno we frazie nominalnej jak i werbalnej w językach polskim i litewskim (por. Roszko, R., 2004; Roszko, D., 2015).

Nie jest naszym celem szczegółowe referowanie poszczególnych znaczeń kwantyfikacyjnych. Zamierzamy jednak ukazać przydatność języka litewskiego w procesie automatyzacji nanoszenia anotacji semantycznej, w szczególności w ujednoznacznianiu wieloznacznych polskich wykładników. Podkreślmy, niedopowiedzenie kwantyfikacyjne jest bardzo rozpowszechnione w języku polskim, dlatego odwołanie się do przejrzystych formalnie litewskich jednoznacznych wykładników znaczeń kwantyfikacyjnych okazuje się bardzo pomocne nie tylko w opisie samego języka polskiego, lecz również w procesie tworzenia algorytmów na potrzeby przekładu maszynowego.

Analiza danych korpusowych (EKorpPL-LT) ujawnia między innymi taką polsko-litewską zależność: polskim zaimkom z cząstką *-ś* odpowiadają litewskie zaimki albo z cząstką *kaž-* albo z cząstką *nors*, por.:

Pol. Ale potrzebne są *jakieś* na to świadectwa …
Lit. Bet juk reikia *kokių nors* įrodymų …

Pol. Bezdomnemu przydarzyło się *coś*, co można porównać jedynie do paraliżu.
Lit. Benamį ištiko *kažkas* panašaus į paralyžių.

Szczegółowa analiza płaszczyzny semantycznej tego typu zdań dostarcza następujących wniosków. Polskim zaimkom z cząstką *-ś* użytym w znaczeniu egzystencjalnym właściwym odpowiadają litewskie zaimki z cząstką *kaž-*.

Natomiast polskim zaimkom z cząstką *-ś* użytym w znaczeniu ogólnym zwyczajowym/ograniczonym odpowiadają litewskie zaimki z cząstką *nors*[5]. Zaobserwowany i przytoczony wyżej fakt pozwala zautomatyzować opis semantyczny wieloznacznych jednostek (tj. jednostek o niedopowiedzianej kwantyfikacji) we wszystkich wielojęzycznych korpusach, w których jednym z języków jest właśnie język litewski.

Warto w tym miejscu przytoczyć przykład na możliwość wykorzystania litewskich jednoznacznych wykładników kwantyfikacji zakresowej również w grupie werbalnej w ujednoznacznianiu polskich odpowiedników, por.:

Pol. Od wczesnego rana świeciły jego siwiejące włosy i niebieskie oczy.
Lit. Tad nuo ankstyvo ryto švies*davo* jo žilstantys plaukai ir mėlynos akys.

Zawarta w powyższym zdaniu polskim forma werbalna świeciły jest wieloznaczna. Jednak w oparciu o litewską formę švies*davo*, z charakterystycznym sufiksem *-dav-*, jesteśmy w stanie jednoznacznie określić typ kwantyfikacji i znaczenie. Mianowicie – kwantyfikacja ogólna i znaczenie ogólne zwyczajowe. Oczywiście niekiedy w polskim tekście znaczenia ogólne zwyczajowe zostają wyeksponowane, por.:

Pol. – Mówił *bywało*: „Krysiu, poczekaj tylko! Jak wukonomem mnie zrobią, ożenię się z tobą."
Lit. – *Saky**davo*: „Ule, palūkėk tiktai! Kai padės mane urėdu, vestuves kelsiva!".

3.2 Modalność możliwościowa

Cechą charakterystyczną zdań modalnych możliwościowych jest obecność funktora możliwości. Poniżej zostaną przedstawione dwa typy modalności możliwościowej: hipotetyczna i imperceptywna. Typowym wykładnikiem znaczeń obu kategorii w języku polskim są leksemy. W języku litewskim obok leksemów występują również regularne wykładniki morfologiczne (formy tzw. trybu modus relativus). Obecność tychże powoduje, że w tekście litewskim można jednoznacznie określić granice między tekstem nacechowanym modalnie a tekstem nienacechowanym, por.:

Pol. *Podobno* przyjechał z rodziną pod wieczór. Spotkał się z burmistrzem nad morzem.
Lit. Tas su šeima *atvažiavęs* vakare. Jis *susitikęs* su meru prie jūros.

5 Szerzej o poszczególnych znaczeniach kwantyfikacyjnych (por. Roszko, D., 2015).

W polskim tekście leksem *podobno* – wykładnik znaczeń imperceptywnych – pojawia się tylko w zdaniu pierwszym. W litewskim wariancie tekstu nacechowanie imperceptywne jest obecne w obu zdaniach. Dlatego już sam ten fakt można wykorzystać do wzbogacenia opisu polskich jednostek. W danym wypadku można polskiej formie *spotkał się* przypisać wartość imperceptywną. Być może dla przeciętnego użytkownika języka polskiego będzie to niewiele wnosząca informacja, jednak kiedy trzeba to polskie zdanie przetłumaczyć na język bułgarski, wówczas informacja ta będzie niezwykle przydatna w wyborze tzw. formy nieświadka w języku bułgarskim.

3.2.1 Modalność hipotetyczna

Jest to kategoria zdaniowa służąca wyrażeniu subiektywnego stosunku nadawcy do wypowiadanych przez siebie sądów (Maldžieva, 2003). Maldžieva (2003), podobnie D. Roszko (2015), wyróżnia 6 poziomów stopnia prawdopodobieństwa. Szerzej o samej kategorii (por. Maldžieva, 2003) oraz o wykładnikach w językach polskich i litewskich (por. Roszko, D., 2015).

Przyjrzyjmy się poniższym zdaniom:

Pol. – *Musiałeś* go gdzieś *zostawić* – rzekł Kubuś Puchatek.
– Ktoś *musiał* mi go *zabrać* – powiedział Kłapouchy. –
I jak tu mieć dla nich serce? – dodał po dłuższej chwili milczenia.
Lit. – *Būsi* ją kur nors *palikęs*, tarė Pūkuotukas.
– Kas nors *bus pasiėmęs*, – pasakė Nulėpausis. – Va kokie,
– pridūrė ilgokai patylėjęs.

Polskim wieloznacznym konstrukcjom *musi* + *bezokolicznik* (*musiałeś zostawić*, *musiał zabrać*) odpowiadają jednoznaczne litewskie konstrukcje morfologiczne, służące wyrażeniu znaczeń hipotetycznych (*b*ūsi *palikęs*, *bus pasiėmęs*). Zatem, tak jak w wypadku znaczeń kwantyfikacji zakresowej, również i tu można zastosować projekcję znaczeń hipotetycznych odczytywanych z jednoznacznych litewskich wykładników na wieloznaczne polskie ekwiwalenty.

Warto również w wyżej przytoczonych zdaniach zwrócić uwagę na formy zaimkowe i przysłówkowe: wieloznaczne pol. *gdzieś*, *ktoś* i jednoznaczne lit. *kur nors*, *kas nors* (por. wyżej p. 3.1.).

Z przedstawionej przez D. Roszko (2015: 246) analizy zasobów korpusowych wynika również, że choć liczba leksykalnych wykładników hipotetyczności okazała się zdecydowanie wyższa w języku polskim niż w języku

litewskim[6], to jednak różnorodność ich użycia w tekście zdecydowanie przemawia na korzyść języka litewskiego. O ile w wypadku języka polskiego można mówić o wyraźnie dominujących wykładnikach-przedstawicielach swoich grup (por. pol. *chyba* obejmujące 95% użyć wszystkich wykładników w grupie H4, pol. *na pewno* z 78% użyć w ramach grupy H6 czy pol. *może* z 49% użyć w ramach grupy H5), o tyle w języku litewskim – już nie. Litewski wykładnik o najwyższej częstotliwości użycia w ramach swojej grupy charakteryzuje wielkość 35% (lit. *gal*, grupa H5), kolejne zaś to już 17% (lit. žinoma, również należący do grupy H5) i 14% (lit. *tikriausiai*, grupa H6).

3.2.2 Modalność imperceptywna

Jest to kategoria zdaniowa służąca wyrażeniu subiektywnego stosunku aktualnego nadawcy do powtórnie wypowiadanych treści, (por. Korytkowska, 1978; Korytkowska, Roszko, R., 1997). Korytkowska, D. Roszko oraz R. Roszko wyróżniają 2 poziomy stopnia prawdopodobieństwa (neutralny i wzmocniony), (por. Korytkowska, Roszko, R., 1997; Roszko, R., 1993; Roszko, D., 2015). Szerzej o samej kategorii (por. Korytkowska, Roszko, R., 1997) oraz o wykładnikach w językach polskim i litewskim (por. Roszko, R., 1993; Roszko, D., 2015).

O tym, że do wyrażenia treści imperceptywnych dochodzi zdecydowanie rzadziej w języku polskim niż w litewskim świadczą chociażby dysproporcje w użyciu polskich i litewskich leksemów-wykładników imperceptywności zarejestrowane w EKorpPL-LT, por.:

Pol. Kiedy – przestraszony sztuczkami Korowiowa, który ukazał mu kota, trzymającego na widelcu marynowany grzyb – stracił przytomność w mieszkaniu wdowy po jubilerze, leżał tam, dopóki Korowiow, natrząsając się zeń, nie wcisnął mu na głowę wojłokowego kapelusza i nie wysłał go na moskiewskie lotnisko, uprzednio zasugerowawszy oczekującym tam na Stiopę przedstawicielom wydziału śledczego, że Stiopa *wysiądzie* z samolotu, który przyleciał z Sewastopola.

Lit. Apalpęs juvelyro našlės bute, kur buvo išgąsdintas Korovjovo triuko su katinu, pasimovusiu ant šakutės marinuotą grybą, jis pragulėjo tame bute tol, kol Korovjovas tyčiodamasis užmaukšlino jam ant galvos veltinę skrybėlę ir nudangino jį į Maskvos aerouostą, pirma dar įteigęs Stiopą sutinkantiems kriminalinės paieškos atstovams, kad Stiopa *neva išlipęs* iš lėktuvo, atskridusio iš Sevastopolio.

W litewskim wariancie obok wykładnika leksykalnego *neva* pojawia się wykładnik morfologiczny *išlipęs*. W polskim tekście nie ma żadnego wykładnika

6 Polskich 88 do 72 litewskich. W podanych liczbach uwzględniono tylko te wykładniki, które zarejestrowano przynajmniej dziesięciokrotnie.

wskazującego na treści imperceptywne. Te treści, jak widać, pozostają niewyrażone w języku polskim. Z analizy odrzuconych tekstów z EKorpPL-LT wynika, że przeciętny polski tłumacz języka litewskiego oddałby litewskie …*kad Stiopa neva išlipęs iš lėktuvo*… polskim *…że <u>podobno</u> *Stiopa wysiądzie z samolotu*…, natomiast tłumacząc polskie …że Stiopa wysiądzie z samolotu… – litewskim … *kad Stiopa išlips iš lėktuvo*… W obu wypadkach stwierdza się podążanie tłumacza za formą oraz brak refleksji nad semantyczną strukturą zdania. W pierwszym wypadku pod wpływem podwojonego wykładnika znaczeń imperceptywnych w języku litewskim, te zostałyby „przemycone" do języka polskiego. W drugim (odwrotnym) wypadku – znaczenia imperceptywne zostałyby „wyrugowane" z treści zdania litewskiego. Podobny proces „typowych" przekładów między językami polskim i bułgarskim (w którym istnieje morfologiczny wykładnik znaczeń modalności imperceptywnej) stwierdzają M. Korytkowska i R. Roszko (Koseska, Korytkowska, Roszko, 2007).

Analiza ekwiwalentnych zdań polskich i litewskich (zawierających tylko morfologiczny wykładnik imperceptywności) ujawnia kolejną zależność. Jest nią brak jakiegokolwiek wykładnika imperceptywności w języku polskim, jeśli w zdaniu litewskim zostaje zastosowany morfologiczny wykładnik niewzmocnionych znaczeń imperceptywnych, por.:

Pol.	Uderzenie było tak mocne, że pojazd dosłownie wjechał pod ciężarówkę.
Lit.	Smūgis *buvęs* toks stiprus, kad automobilis tiesiogine pra sme palindo po sunkvežimiu.

Litewskie *buvęs* – to morfologiczny wykładnik znaczeń imperceptywnych.

Z kolei, gdy w języku litewskim zostaje zastosowany wykładnik wzmocnionych znaczeń imperceptywnych, to wówczas w języku polskim stwierdza się użycie imperceptywnego wykładnika leksykalnego, ewentualnie leksykalnego sprzężonego z wieloznaczną konstrukcją paramorfologiczną *ma* + *bezokolicznik*, por.:

Pol.	Jan powiedział, *jakoby* brat *miał się zatrzymać* u ciotki.
Lit.	Jonas pranešė, kad tasai *esąs apsigyvenęs* pas tetą.

Anotacja semantyczna w wypadku wieloznacznych polskich wykładników okazuje się ułatwiona, gdy zestawimy konkretne odpowiadające sobie zdania polskie i litewskie, por.:

Pol.	Miał przyjechać.
Lit.	Jis turėjo atvažiuoti.

oraz

Pol. Miał przyjechać.
Lit. Jis *atvažiavęs*.

Dla pierwszej pary zdań, w oparciu o postać litewską, stwierdza się brak znaczeń imperceptywnych, tym samym wieloznaczna polska konstrukcja *ma + bezokolicznik* nie jest w danym wypadku wykładnikiem znaczeń imperceptywnych. Natomiast w parze drugiej litewski jednoznaczny morfologiczny wykładnik niewzmocnionych znaczeń imperceptywnych *atvažiavęs* wskazuje na imperceptywne nacechowanie polskiej konstrukcji *ma + bezokolicznik*

4. Podsumowanie

Zachowawczy charakter języka litewskiego przyczynia się do przejrzystości struktur formalnych oraz powiązań między formą a jej funkcją. W języku polskim (szerzej słowiańskim) wprowadzony na pewnym etapie rozwoju języka element funkcjonalny nierzadko w wyniku zmian fonetycznych zaciera swą postać, ta zaś po latach zostaje uwolniona od pierwotnego znaczenia nierzadko nim dojdzie do pełnego jego (znaczenia) zgramatykalizowania. Zachodzące więc w językach słowiańskich zmiany fonetyczne, w tym także w języku polskim, sprawiają, że pewne wartości semantyczne nie są ujawnione na poziomie formalnym, por. pol. *Niech minister się schowa*. i dwa możliwe warianty litewskie: *Tegul ministras* (r.m.) *nesilygina*. i *Tegul ministrė* (r.ż.) *nesilygina*., w których zauważamy rozróżnienie osób płci żeńskiej i męskiej. Podobne „niedoskonałości" polszczyzny można wskazać na przykładzie form deminutatywnych. W wyniku ograniczeń formalnych pewne polskie formy imienne nie posiadają form deminutatywnych, lub jeśli je tworzą, to z pewnością nie są one stylistycznie neutralne, por. pol. Polska i *--- oraz lit. *Lietuva* `Litwa' i *Lietuvėlė* (zdrobnienie od *Lietuva* 'Litwa'). Takie przykłady można mnożyć, por. jeszcze jeden – litewskie *auti* i polskie już dzisiaj raczej sporadyczne *obuwać*, w którym w odniesieniu do litewskiej formy obserwujemy zarówno proces perfektywizacji i wtórnej imperfektywizacji. O odpowiedniku fonetycznym litewskiego dyftongu *au* pisaliśmy w punkcie 2. – jest nim w językach słowiańskich *u*. Zatem prosty odpowiednik lit. *auti* w języku polskim mógłby mieć nadal postać **uć*, przybrał jednak postać niewyobrażalnie złożoną *ob-u̲-wa-ć̲*.

Kwantyfikacja zakresowa jest istotnym elementem semantycznej struktury zdania. Dlatego w obliczu charakterystycznego dla języka polskiego niedopowiedzenia kwantyfikacyjnego zestawienie tekstów polskich z litewskimi pozwala tę wieloznaczność usunąć. Podobny efekt ujednoznacznienia form polskich można osiągnąć w zakresie modalności możliwościowej, zwłaszcza przedstawionej tu imperceptywności.

Prezentowana tu idea polegająca na automatyzacji opisu funkcji polskich form w oparciu o jednoznaczne litewskie wykładniki z zakresu kwantyfikacji zakresowej czy modalności możliwościowej jest w naszym rozumieniu jedynie wstępem do przyszłego nieuniknionego procesu łączenia nadal rozproszonych zasobów korpusowych do postaci korpusów wielojęzycznych, w których przedstawiona tu w przykładach anotacja znaczeń (anotacja opisująca nie morfologiczne parametry formy, lecz jej aktualne funkcje wynikające z użycia) może zostać zastosowana z pożytkiem nie tylko dla pełniejszego opisu poszczególnych języków czy badań wybitnie językoznawczych (np. opisowych, kontrastywnych), lecz również dla ustalania algorytmów międzyjęzykowej ekwiwalencji ze wszelkimi tego konsekwencjami dla wszelkich nauk stosowanych bazujących na języku (takich jak proces nauczania języka, przekład tradycyjny czy maszynowy i in.).

Nie zakładamy, że to właśnie język litewski ma być podstawą do automatyzacji procesu anotacji semantycznej. Każdy bowiem język posiada pewne jednoznaczne wykładniki określonych znaczeń, które można dołączać do uniwersalnej już wówczas podstawy. Opierając się na jednoznacznych wykładnikach w jednym języku można zawęzić materiał w drugim języku do zgodnego z założonymi parametrami (wyznaczonymi przez jednoznaczne wykładniki języka wyjściowego) i w ramach tak wyselekcjonowanego materiału szukać reguł, które – niezauważane przy tradycyjnym podejściu do języka – mogą zostać ujawnione.

Proponowane tu podejście opisu znaczeń (funkcji) poszczególnych form w wielojęzycznych korpusach wydaje się nieuniknioną przyszłością wielojęzycznych zasobów. Jak już niejednokrotnie w literaturze przedmiotu było podnoszone, tym, co łączy języki nie są formy (i ich własności gramatyczne), lecz płaszczyzna znaczeniowa – inaczej funkcje form (por. np. Weinsberg, 1983).

Literatura

Ajdukiewicz, Kazimierz (1965): *Logika pragmatyczna*. Warszawa: Państwowe Wydawnictwo Naukowe.

Karolak, Stanisław (2001): *Od semantyki do gramatyki*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Korytkowska, Małgorzata (1978): Ze studiów nad modalnością w języku bułgarskim. *Studia z Filologii Polskiej i Słowiańskiej* XVII, 263–288.

Korytkowska, Małgorzata, Roszko, Roman (1997): *Gramatyka konfrontatywna bułgarsko-polska*, tom 6, część 2. *Modalność imperceptywna.* Warszawa: Slawistyczny Ośrodek Wydawniczy.

Koseska-Toszewa, Violetta (1982): *Semantyczne aspekty kategorii określoności/nieokreśloności (na materiale z języka bułgarskiego, polskiego i rosyjskiego).* Wrocław: Zakład Narodowy im. Ossolińskich.

Koseska-Toszewa, Violetta (2013): About Certain Semantic Annotation in Parallel Corpora. *Cognitive Studies | Études cognitives* 13, 67–78. DOI: 10.11649/cs.2013.004.

Koseska-Toszewa, Violetta, Korytkowska, Małgorzata, & Roszko, Roman (2007): *Polsko-bułgarska gramatyka konfrontatywna.* Warszawa: Wydawnictwo Akademickie „Dialog".

Koseska-Toszewa, Violetta & Roszko, Roman (2015): On Semantic Annotation in Clarin-PL Parallel Corpora. *Cognitive Studies | Études cognitives* 15, 211–236; DOI: 10.11649/cs.2015.016.

Maldžieva, Vjara (2003): *Gramatyka konfrontatywna bułgarsko-polska*, tom 6, część 3. *Modalność*: *hipotetyczność*, *irrealność*, *optatywność i imperatywność, warunkowość).* –Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, Danuta (2006): *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, Danuta (2015): *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (na tle literackich języków polskiego i litewskiego).* Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, Danuta & Roszko, Roman (2013): Experimental Polish-Lithuanian Corpus with the Semantic Annotation Elements. *Cognitive Studies | Études cognitives* 13, 97–111; DOI: 10.11649/cs.2013.006

Roszko, Roman (1993): *Wykładniki modalności imperceptywnej w języku polskim i litewskim*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Roszko, Roman (2004): *Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim)*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Weinsberg, Adam (1983): *Językoznawstwo ogólne*. Warszawa: Państwowe Wydawnictwo Naukowe.

Косеска-Тошева, Виолетта & Гаргов, Георги (1990): Българско-полска съпоставителна граматика. (том 2. Семантичната категория определеност/неопределеност). – София.