Wydział Informatyki i Zarządzania
Katedra Inteligencji Obliczeniowej
Politechnika Wrocławska

Marta Dobrowolska-Pigoń, Agnieszka Dziob, Barbara Nowakowska, Justyna Wieczorek

Procedura korekty Słowosieci

Ver. 1

Słowa kluczowe:
wordnet, plWordNet, Słowosieć, relacje leksykalne, leksykografia, leksykologia, semantyka, język polski, diagnostyka

WROCŁAW 2020

Spis treści

Informacje wstępne

Niniejszy dokument ma na celu przedstawienie propozycji ręcznej, merytorycznej korekty Słowosieci przez lingwistów polskich. System korekty automatycznej i półautomatycznej został przedstawiony w osobnym dokumencie (zob. poz. 3. poniżej). W tej Procedurze zostały zebrane częste szczegółowe problemy wiążące się z pracą lingwistów, które dają się uogólnić i wymagają rozstrzygnięcia. Nie zajmujemy się w nim problemami jednostkowymi. Zakładamy, że prezentowana, pierwsza wersja Procedury będzie rozszerzana w miarę postępów prac związanych z redakcją i korektą Słowosieci.

Dokumenty powiązane

Pewne elementy opisu w Słowosieci są skodyfikowane w innych dokumentach:

1. Obowiązujący dokument dotyczący struktury komentarza i definicji w Słowosieci: https://docs.google.com/document/d/1SK0AtQ81GSQTj_0RHUj5l0o4z0-jcYuPTddfBCgiiqE/edit
2. Dokument z opisem systemu kwalifikatorów w Słowosieci: https://docs.google.com/document/d/1NIFcGjW33RlyQCmPvLYLyQhwUFgObISpjgrPriDz2oA/edit#heading=h.e4ks0qqgpt5g
3. Dokument z opisem wstępnej diagnostyki automatycznej: https://docs.google.com/document/d/1Dv5llRDdCNBee-kByln8CBTnTPqTc0esq1DJmslTLqs/edit?usp=sharing
4. Dokument z opisem relacji leksykalno-semantycznych dla czasownika w Słowosieci 4.0, w którego rozdziale Procedura zawarto najważniejsze informacje dotyczące posługiwania się systemem redakcji i korekty w Słowosieci: https://docs.google.com/document/d/1SK0AtQ81GSQTj_0RHUj5l0o4z0-jcYuPTddfBCgiiqE/edit#

System flag

Od wersji 3.1 Słowosieci lingwistom zostało udostępnione pole z etykietami („flagami") służącymi do oceny poprawności jednostek i synsetów w aplikacji Wordnet Loom Editor. Domyślnie każda jednostka lub synset ma ustawiony status *Nieprzetworzony*, a lingwista ma możliwość ręcznej zmiany statusu. Do wyboru są następujące statusy:

1. *Błąd* – poważny błąd, który jest zbyt czasochłonny, by można go w danej chwili poprawić;
2. *Nowy* – jednostka nowo wprowadzona (zakładamy, że jest wprowadzona dobrze, ale może ulec weryfikacji/zmianie);
3. *Sprawdzony* – informacja o tym, że lingwista zweryfikował dany lemat/synset;
4. *Znaczenie* – informacja o tym, że lingwista wprowadził jedynie dane znaczenie dla danego lematu;

5. *Częściowo przetworzony* - informacja o tym, że lingwista zweryfikował komplet znaczeń danego lematu, ale nie zajmował się relacjami (hiperonimią, bliskoznacznością itp.).

W przypadku wyboru oznaczenia *Błąd* w oknie komentarza pojawi się dodatkowe okno, w którym należy skrótowo opisać to, dlaczego dana jednostka lub synset zostały uznane za błędne. Okno z miejscem na opis pojawi się także przy wyborze flagi *Częściowo przetworzony*. Można w nim wpisywać, co zostało zrobione, a co nie, dlatego też flaga ta może mieć również inne zastosowania, niż to, które zostało wymienione w punkcie 5.

Sprawdzony

W Słowosieci lingwiści pracują z lematem, uzupełniając mu komplet znaczeń, a wprowadzanie pojedynczych znaczeń dopuszcza się tylko w przypadkach, kiedy są one niezbędne do opisu innej jednostki, nad którą lingwista pracuje. Dopiero jednostki z kompletu mogą uzyskać status *Sprawdzony*. W przypadku synsetów *Sprawdzony* mogą uzyskać te z nich, w których wszystkie jednostki są *Sprawdzone*.

Kilka zasad technicznych

1) W komentarzu do jednostki piszemy uwagi dotyczące elementów komentarza (kwalifikatora, glosy, przykładu użycia, linku), znaczenia jednostki bądź kompletu znaczeń w lemacie, relacji jednostki, anotacji emotywnej jednostki bądź kolokacji;
2) W komentarzu synsetu piszemy uwagi odnośnie do relacji synsetów oraz rzutowania na inne leksykony (np. Princeton WordNet lub Walentego), a także składu synetu;
3) Dodajemy komentarz do jednostki lub synsetu, do których mamy uwagi, a nie do tych, które są z nimi powiązane relacjami;
4) Komentarz musi określać, rodzaj błędu, na jaki wskazujemy, komentarz musi być jednoznaczny w bazie danych w odniesieniu do jednostki lub sysnetu, a nie kontekstowo powiązany z siatką relacji, w jakiej ta jednostka lub synset występują w wizualizacji.

Jakie błędy poprawiamy od razu

1) Literówki w kwalifikatorach, glosach, przykładach użycia;
2) Brak kwalifikatora, jeśli reszta komentarza nie jest pusta;
3) Błędy gramatyczne w glosach i przykładach użycia, także użycie liczby mnogiej do definiowania słowa w liczbie pojedynczej;
4) Braki definiowanego słowa w przykładzie użycia.

Wstępne uwagi merytoryczne

Polisemia

Słowosieć jest słownikiem relacyjnym, w związku z czym to relacje różnicują znaczenia. Jeśli dwa znaczenia zostały wyróżnione przez lingwistę i opisane w postaci odrębnych jednostek

leksykalnych, jednak zestawy ich relacji nie różnią się (np. są kohiponimami bez żadnych dodatkowych relacji poza hiponimią), z punktu widzenia semantyki relacyjnej mamy do czynienia z tym samym znaczeniem.

W artykule The chicken and egg problem… zostały opisane relacje konstytutywne w Słowosieci. Relacje konstytutywne w wordnetach definiuje się jako relacje częste, powtarzalne w różnych wordnetach i możliwe do wykorzystania w różnorodnych zastosowaniach NLP. W momencie powstania tego artykułu za relacje konstytutywne uznano wszystkie relacje synsetów poza bliskoznacznością i fuzzynimią. Od tego czasu zestaw relacji uległ rozszerzeniu, uznajemy więc, że oprócz dwóch wymienionych relacjami konstytutywnymi nie są: określnik (dla przymiotnika), subiekt i obiekt (dla czasownika i rzeczownika) oraz nowe relacje rzeczownikowe, zdefiniowane na podstawie zleksykalizownych relacji paradygmatycznych: subiekt przy niewyrażonym predykacie, wytwór / rezultat przy niewyrażonym predykacie, miejsce przy niewyrażonym predykacie oraz czas przy niewyrażonym predykacie.

Relacje konstytutywne nie są równoznacznie z relacjami wymaganymi (zob. Relacje wymagane). W przypadku relacji wymaganych, niebędących konstytutywnymi, jednostka docelowa dla relacji dziedziczy siatkę relacji synsetów, w tym relacje konstytutywne, po jednostce źródłowej. Tak się dzieje np. w przypadku bliskoznaczności czy żeńskości.

Definicja relacji konstytutywnych została przede wszystkim powołana na potrzeby łączenia wordnetów oraz zastosowań NLP. Nie zawsze oddaje stan zgodny z wiedzą o języku. W związku z tym przyjęliśmy, że relacje derywacyjne również różnicują znaczenia w tym sensie, że jedna jednostka leksykalna nie może pochodzić od dwóch podstaw słowotwórczych. Wyjątkiem są sytuacje, w których pochodzenie derywacyjne jest niepewne - wtedy istnieje możliwość dodania więcej niż jednej relacji derywacyjnej w kierunku podstawy słowotwórczej.

Antropocentryzm

Reguła antropocentryzmu była stosowana przy opisie znaczeń w Słowosieci 2.0, zwłaszcza w przypadku przymiotnika, w którym jednym z kryteriów wyróżniania znaczeń był podział na określenia człowieka i „inne”. Jest ona również widoczna przy wydzielaniu znaczeń dla czasownika z tego okresu, a także przy opisie znaczeń niektórych rzeczowników - zwłaszcza takich, które mogą mieć bardzo szerokie znaczenie i opisują nosicieli cech (np. *brudas*) oraz takich, które mogą nazywać osobę lub instytucję (np. *ubezpieczyciel*).

Należy pamiętać, że głównym elementem opisu znaczenia w Słowosieci są relacje. Znaczenia opisujące jakąś cechę, podmiot, sytuację, należy więc łączyć z takimi hiperonimami, które traktowałyby dane zjawisko wystarczająco szeroko. Szczególnie trzeba mieć na uwadze całe ścieżki hiperonimiczne, nie tylko bezpośredni hiperonim.

Zbyt szerokie definicje

Zadaniem definicji jest definiowanie, a problemem zbyt szerokich definicji to, że nie spełniają one swoich funkcji. W glosach, które nie definiują znaczenia, bardzo często użyte są zaimki nieokreślone w przypadkach, w których można zastosować konkretne słowa. Przykład

glosy, która nie definiuje znaczenia to '*robić coś, działać, skupiając się na określonym zadaniu*' dla *pracować 4* - element 'skupiać się' nie jest cechą dystynktywną znaczenia, a 'robić coś' niewiele o nim mówi (por. *pracować 1 'robić coś, co ma doprowadzić do zrealizowania jakiegoś celu, spełnienia jakiegoś obowiązku'*, w przypadku którego realizacja celu jest cechą dystynktywną znaczenia, wskazywaną przez zaimek 'coś' ze zdania nadrzędnego w glosie).

Zbyt wąskie definicje

Problemem odwrotnym do zbyt szerokich definicji, ale również wiążącym się z rozróżnianiem znaczeń polisemicznych, jest zbyt wąskie definiowanie. Unikamy dawania w glosie jednostki odniesień do tych aspektów jednostki, które nie wiążą się z jej znaczeniem semantycznym, a zawężają je. Zwłaszcza łatwo jest odnosić się do ról semantycznych, oddających nie restrykcje (ani nawet nie preferencje) selekcyjne, ale częste konkordancje (kontekst) czy kolokacje.

Przykład zbyt wąskiej definicji ze Słowosieci 4.0 to *wciągnąć 1,* które ma glosę: *'sprawić, że w specjalnym spisie, na liście lub w bazie danych znajdzie się informacja, że ktoś, coś spełnia określone kryteria'*. Komentarz zwraca uwagę na to, że „specjalność spisu" i „określone kryteria" stanowią zawężenie definicji. Inny przykład to: *nabrać 6* definiowane *'przez jamę ustną przyjąć do organizmu jakąś substancję, zazwyczaj lotną lub płynną'*. Komentarz zwraca uwagę, że w ten sposób nabieramy np. wody do bukłaków, więc dopełnienie „przez jamę ustną" stanowi zbytnie zawężenie definicji.

Aby poprawnie zdefiniować znaczenie dla czasowników, warto się zastanowić, czy wszystkie argumenty, wiążące się z predykatem w glosie, są konieczne, tzn. służą definicji. W przypadku przymiotników czy przysłówków warto pomyśleć, czy nie próbujemy definiować znaczenia za pomocą kolokatów. Należy się też zastanowić, czy staramy się wykorzystać w glosie hiponimy, meronimy albo kohiponimy definiowanej jednostki. Nie możemy tego zrobić, jeśli nie uda nam się przedstawić zamkniętego ciągu wyliczenia, który jest równoznaczny definiowanemu znaczeniu. W przypadku kohiponimów znalezienie takiego ciągu jest, z logicznego punktu widzenia, niemożliwe. Unikamy więc definiowania za pomocą kohiponimów.

Synonimia

Synonimia jest w Słowosieci (i innych wordnetach) relacją wtórną, której istnienie wynika z tego, że dwie jednostki reprezentujące ten sam lemat mają dokładnie takie same relacje. Nie jest więc możliwe, żeby kohiponimy różniły się tylko glosami, bez istnienia żadnych innych relacji różnicujących znaczenia. Z perspektywy semantyki relacyjnej są one synonimami.

Z formalnego punktu widzenia synonimia nie jest też relacją zdefiniowaną w bazie SQL. Synonimy da się jedynie wyszukać na podstawie składu synsetu.

W Princeton WordNet jako synonimy określa się słowa (jednostki leksykalne) zastępowalne w kontekście. Tak opisywane synonimy mają wspólną glosę przypisaną do synsetu i wspólny przykład użycia. W Słowosieci stosuje się kwalifikatory, glosy i przykłady użycia w odniesieniu do jednostki. O ile glosa, oddająca znaczenie leksykalne, powinna być w przypadku

synonimów wspólna (na poziomie opisu jednostki może to się przejawiać dokładnie takimi samymi glosami dla znaczeń lub glosami definiującymi dokładnie to samo znaczenie), to jednostki synonimiczne mogą różnić się rejestrem (w ramach rejestrów podobnych, zob. Kwalifikatorów system) oraz przykładami użycia. To drugie wynika z faktu, że nie uznajemy walencji za czynnik decydujący o różnicy znaczeniowej, w związku z czym mogą zdarzyć się przypadki, w których dwie jednostki leksykalne nie będą dokładnie zastępowalne w kontekście. Na przykład jednostki *podpisać* [coś] oraz *podpisać się* [pod czymś] z naszego punktu widzenia reprezentują to samo znaczenie.

Warto również zwrócić uwagę na definiowanie jednostek wielowyrazowych z przyimkiem w lemacie w Princeton WordNet. Nie wszystkie z nich uznalibyśmy za jednostki wielowyrazowe w Słowosieci (zob. Wielowyrazowe jednostki leksykalne) - często przyimek wskazuje na rekcję danego czasownika.

Relacje wymagane

Do poprawnego zdefiniowania synsetu, konieczne jest opisanie go przynajmniej za pomocą którejś z poniższych relacji:

1) hiponimii w przypadku wszystkich części mowy,
2) typu w przypadku nazwy własnej,
3) własności cechy w przypadku przymiotnika i przysłówka,
4) odwrotnej mero- i holonimii w przypadku rzeczowników,
5) bliskoznaczności w przypadku słów nacechowanych, należących do wszystkich części mowy.

Ponadto istnieją relacje jednostek, które wystarczają do prawidłowego opisania jednostek w synsetach, pod warunkiem, że każda z jednostek wchodzących w skład synsetu będzie miała którąś z tych relacji. W przeciwnym razie konieczne jest wprowadzenie którejś z powyższych relacji synsetów. Do zestawu relacji jednostek, o których mowa, należą:

1) nacechowanie (deminutywność oraz ekspresywność i augumentatywność) i żeńskość dla rzeczowników,
2) synonimia międzyparadygmatyczna dla przymiotników relacyjnych lub derywacyjność dla przymiotników relacyjnych pochodzących od więcej niż jedna podstaw słowotwórczych,
3) stopień wyższy lub najwyższy dla przymiotników lub przysłówków w tym stopniu.

Przyjmujemy, że relacji wymaganych nie muszą mieć rzeczowniki z najwyższego poziomu hierarchii (*bhp*) oraz czasownikowe synsety sztuczne stojące na górze hierarchii dla czasownika.

W przypadku bliskoznaczności synset nacechowany „dziedziczy” relacje po swoim nienacechowanym odpowiedniku, jest więc wpisany w hierarchię Słowosieci za pośrednictwem jego relacji. To samo dotyczy nacechowanych jednostek i derywatów, które posiadają jedynie relacje jednostek wymagane do ich zdefiniowania. W związku z tym nie jest konieczne tworzenie osobnych drzew hierarchicznych dla feminatywów lub jednostek ekspresywnych (zob. Rzeczowniki) . Można wpisać dany synset z jednostkami nacechowanymi w takie drzewo, jeśli ono już istnieje, ale nie jest to konieczne do jego poprawnego zdefiniowania.

Elementy komentarza

Znaki interpunkcyjne i podział wiersza

Średnik

W dokumencie „Komentarze w Słowosieci” jest dozwolone użycie średnika jednokrotnie w glosie.

Średnik pełni funkcję separatora tabel w bazach danych SQL, czyli w formacie bazodanowym Słowosieci. Jest to widoczne np. przy zapisie danych z bazy w tzw. formacie wertykalnym (np. csv) - tam, gdzie są średniki, dane są dzielone, tzn. pojawiają się w oddzielnych kolumnach arkusza. Większość narzędzi informatycznych, odwołujących się do bazy Słowosieci, nauczyła się radzić sobie ze średnikiem w glosach. Błąd związany z dzieleniem tabel w miejscu średnika pojawia się natomiast przy wydobywaniu danych ze Słowosieci na potrzeby lingwistów i użytkowników.

Z punktu widzenia informatyki kropka jest lepsza niż średnik, ale średnik przyjął się w tradycji leksykograficznej - za nim zwykło się pisać informacje metalingwistyczne.

Co robić ze średnikiem? Najlepiej go nie używać, a w miejscu, w którym widzimy średnik, który można zamienić na coś innego bez straty informacji - zmienić. Jeśli już musimy wstawić średnik w glosie, to w dalszym ciągu dopuszczamy jeden średnik na glosę, jednak im mniej średników - tym lepiej dla przejrzystości informacji ze Słowosieci. Wykluczamy natomiast średnik w komentarzach. Lepiej stawiajmy kropkę.

Podział wiersza

Podobnie format wertykalny staje się nieczytelny przy twardej spacji, która wygląda dobrze tylko w WordNetLoomie. Unikamy jej, podobnie jak unikamy pustych linii pomiędzy poszczególnymi elementami komentarza.

Ukośnik

W przeciwieństwie do średnika, kropki oraz twardej spacji, ukośniki nie są traktowane w plikach jako znaki podziału. Ponadto używanie ukośnika jest zakorzenione w opisie leksykograficznym, w przypadku, na przykład, czasowników dwuaspektowych, czy formuł odpowiadających opisom przypadków, w jakich się używa rzeczownika bądź przymiotnika (np. „ktoś/coś”, „o kimś/o czymś” itd.). Użycie ukośnika uznajemy za uprawnione w wypadku, kiedy wskazuje on na pewne metajęzykowe elementy definicji. Nie należy go jednak nadużywać. Wykluczamy więc stosowanie ukośnika tam, gdzie powinien stać inny znak interpunkcyjny lub wyrażenie omowne (np. przy wyliczeniach lub alternatywach).

Forma glosy

Glosa powinna zawierać *genus proximum* i *differentia specifica*, wyrażone w sposób opisany w dokumencie Komentarze w Słowosieci. W dokumencie tym wykluczyliśmy jednowyrazowe glosy, które nie zawierają tych dwóch elementów, a z punktu widzenia przetwarzania języka naturalnego nie spełniają kryterium przydatności. Jednowyrazowa glosa jest więc błędem, który należy poprawić.

Błędem opisanym we wspomnianym dokumencie jest również definicja zakresowa, w której nie jesteśmy w stanie wskazać zamkniętego zbioru (kończąca się sformułowaniem *itp.* lub podobnym). Podtrzymujemy te ustalenia, a definicje takie będą kwalifikowane jako *Błąd*.

Dokument ten mówi także, że glosa powinna być napisana poprawną polszczyzną. Mimo że dalecy jesteśmy od rygorystycznej poprawności, uznajemy, że język glosy powinien być staranny i poprawny. Powinna być ona napisana polszczyzną ogólną, uznawaną przez skodyfikowaną normę. Miejscem, gdzie możemy użyć polszczyzny z innych rejestrów lub wskazać na nieprototypowy kontekst jednostki, jest przykład użycia. Uznajemy, że glosy są w Słowosieci odwołaniem do tradycyjnej leksykografii i takimi powinny pozostać. Z tego też powodu unikamy w glosie pospolityzmów i potocyzmów i nawet jednostki, które mają kwalifikatory inne niż ogólny staramy się opisać za pomocą glos w polszczyźnie ogólnej, zgodnej z normami.

W glosach unikamy również wartościowania, jeśli wynika ono z pragmatyki, nie semantyki. Na przykład w semantyce przymiotnika *kiepski* zawiera się negatywne wartościowanie, natomiast negatywne wartościowanie rzeczownika *dziwka* w znaczeniu *prostytutka* jest kwestią pragmatyki i jako takie powinno zostać odnotowane (np. za pomocą elementu metajęzykowego „*pejoratywnie o…*").

Glosy w Słowosieci nie muszą, a czasami wręcz nie powinny mieć formy definicji strukturalnych. Podstawą definiowania w Słowosieci są relacje, nie glosy, i to one odzwierciedlają strukturalne związki w obrębie systemu językowego.

Przykład użycia

Kiedy dodajemy przykłady użycia

Powinniśmy je dawać wtedy, kiedy jest to możliwe. W dokumencie Komentarze w Słowosieci jest napisane, że przykład użycia jest wymagany tylko w przypadku jednostek polisemicznych, bo pomaga odróżnić znaczenia. Jest to prawda, jednak równie przydatny jest w przypadku jednostek monosemicznych - pozwala stwierdzić, że takie znaczenie w ogóle istnieje (zob. Lemat istnieje tylko w słownikach).

Przykładu użycia nie dajemy jedynie w przypadkach, kiedy trudno go znaleźć w źródłach lub wymyślić. Należy się wtedy zastanowić, czy dane słowo istnieje faktycznie (a nie tylko w słownikach). Jeśli odpowiedź jest pozytywna, można w tym wypadku zostawić jednostkę bez przykładu użycia, najlepiej dając odpowiedni komentarz (np. „Trudno znaleźć przykład użycia, istnienie jednostki potwierdzone w korpusie"). Taka jednostka może uzyskać status: *Sprawdzony*.

Jeśli jednostka jest monosemiczna oraz ma kwalifikator i link do Wikipedii, może nie mieć definicji i przykładu użycia. Zakładamy, że artykuł w Wikipedii wystarczy, żeby określić, czy dane znaczenie istnieje. Takie jednostki mogą uzyskać status: *Sprawdzony*.

Poza tym staramy się unikać dawania statusu: *Sprawdzony* jednostkom, które nie mają przykładów użycia. Pracując z kompletem znaczeń lub drzewem, warto dopisać takie przykłady. Oprócz tego, że potwierdzają one istnienie znaczenia, mają nieocenioną wartość w przetwarzaniu języka naturalnego (np. przy systemach WSD).

Forma przykładu użycia

To, jak ma być zbudowany przykład użycia, zostało wyjaśnione w dokumencie Komentarze w Słowosieci. Tam też znajduje się lista źródeł, z których możemy korzystać, jeśli sami nie preparujemy przykładów.

Problem, czy dajemy przykład użycia zaczerpnięty ze źródeł, czy preparowany, rozstrzygamy na korzyść tego pierwszego. Jeśli jednak szukanie przykładu w źródłach jest zbyt czasochłonne (co zdarza się zwłaszcza w przypadku słów rzadkich lub rejestrów, które są rzadkie w Słowosieci), lepiej go spreparować.

Warto w przypadku preparowanego przykładu użycia użyć naturalnej kolokacji, która znajduje się w tekstach w korpusie, i wokół niej budować cały przykład użycia. Ważne, żeby w kolokacji słowo istniało w takim znaczeniu, jak definiowane.

Należy zaznaczyć fakt, który mógłby wydawać się truizmem: w przykładzie użycia musi znajdować się jednostka, która jest definiowana (tzn. lemat w znaczeniu, które odpowiada jednostce leksykalnej). Rejestr błędów pokazuje, że zdarzają się przykłady, w których nie pada definiowana jednostka. Formalnie brak lematu w przykładzie można sprawdzić automatycznie, jednak metody automatyczne nie pomogą w sprawdzaniu, czy w przykładzie znalazło się odpowiednie znaczenie. Częściowym rozwiązaniem problemu będzie metoda stosowania naturalnych kolokacji (opisane powyżej).

Jeśli chodzi o wymóg stosowania w przykładach użycia dla definiowanych czasowników formy finitywnej lub bezokolicznika, który został opisany w wytycznych czasownikowych, rezygnujemy z niego. Niekiedy bardziej naturalną formą dla danego czasownika będzie forma gerundialna lub imiesłowowa. Niekiedy wręcz - będzie ona jedyną możliwą. Staramy się jednak dawać formy osobowe czasownika lub bezokoliczniki tam, gdzie jest to możliwe.

Podobne zastrzeżenia dotyczą form czasowników z „się”. Czasami podanie przykładu ze strukturą V+się jest niemożliwe, a niekiedy przykład ten wygląda nienaturalnie. Nie staramy się na siłę podawać czasownika w formie słownikowej.

Przykład użycia może również zawierać więcej niż jedno zdanie, jeśli ma to sens dla definiowania danej jednostki. Czasami wręcz bardzo trudne jest stworzenie jednozdaniowego przykładu użycia (np. w przypadku znaczeń używanych głównie w mowie, w języku potocznym). Staramy się jednak nie rozbudowywać zbytnio przykładu użycia. Najlepsze (i preferowane) jest jedno zdanie.

Dokument Komentarze w Słowosieci mówił o tym, że wykluczamy cytaty z poezji, jednak niektóre znaczenia książkowe nie występują poza nią. W takich przypadkach cytaty z poezji są dopuszczone, jednak należy je zapisywać w formie jak najbardziej linearnej (tzn. bez podziału na wersy).

Przy konstruowaniu przykładu użycia należy trzymać się zasady, że ma on odwzorowywać naturalny i częsty kontekst występowania danej jednostki. Podajemy w nim pewien wzorzec, który wykorzystują różne systemy przetwarzania języka naturalnego.

Szczegółowe problemy opisu

Co opisujemy, a czego nie

Co opisujemy

W Słowosieci opisujemy jednostki leksykalne należące do czterech części mowy: rzeczownik, przymiotnik, czasownik i przysłówek w ich formach podstawowych. Szczegółowe potencjalne problemy, związane z opisem są rozstrzygane poniżej.

Lemat istnieje tylko w słownikach

Może się zdarzyć, że lemat, który dostaliśmy na liście frekwencyjnej, znajduje się tylko w słownikach, nie ma natomiast żadnych poświadczeń w korpusie. W leksykografii polskiej przyjęło się podążanie za autorytetami. Tzn. jeśli leksykograf jest autorytetem w jakiejś dziedzinie i stwierdził, że jakieś słowo istnieje, to bardzo często jest ono przepisywane do kolejnych dzieł leksykograficznych w takim znaczeniu, w jakim pierwotnie ono wystąpiło.

Słowosieć jest oparta na danych korpusowych, dlatego powinniśmy włączać do niej tylko te jednostki, których wystąpienia udało się poświadczyć. Korpus Słowosieci nie jest zrównoważony i z pewnością nie obejmuje całego słownictwa, jednak o tym, czego w nim nie ma, można powiedzieć z dużą dozą prawdopodobieństwa, że jest rzadkie w polszczyźnie ogólnej i niewłączenie tego do opisu w sposób niezauważalny jedynie obniży skuteczność narzędzi, działających na bazie Słowosieci. Słowa takie mogą również należeć do rejestrów słownictwa, które nie są regularnie opisywane w Słowosieci, lub też nie są w ogóle opisywane, np. słownictwo dawne, archaiczne, regionalizmy.

Jeśli lemat (lub znaczenie) istnieje tylko w słownikach i nie ma potwierdzenia w korpusach ani w Internecie - nie wprowadzamy go, jeśli nie ma takiej wyraźnej potrzeby ze strony użytkownika (zdarzały się przypadki takich wymagań - np. podczas pracy nad słownikiem jidysz zespołu z UW). Jeśli taki lemat (lub znaczenie) już jest opisane - nie usuwamy go. Niektóre z nich mają w komentarzu dopisek mówiący o tym, gdzie lingwista ten lemat lub to znaczenie znalazł. Nie usuwamy tych komentarzy, a jeśli zdarzy nam się znaleźć coś, co nie ma poświadczenia w korpusach - warto taki komentarz zostawić.

Takie jednostki będą mogły mieć status: *Sprawdzony*.

Polskie i niepolskie słowa

Słowosieć ma służyć opisowi polskiego słownictwa. Przyjęliśmy zasadę, że dany lemat wprowadzamy tylko wtedy, jeśli jest zakorzeniony w polskim leksykonie. Wyznacznikiem takiego zakorzenienia jest odmiana w polskim paradygmacie. Jednak jeśli słowo lub jednostka wielowyrazowa jest powszechnie używana, tzn. ma wysoką frekwencję w Korpusie Słowosieci (100 wystąpień lub więcej) lub pojawia się w Narodowym Korpusie Języka Polskiego, a nie ma polskiej odmiany, to można ją wprowadzić, jeśli nie kłóci się to z naszym odczuciem, mówiącym, że ta jednostka należy do systemu języka polskiego. Takimi dobrze rozpoznawanymi jednostkami wielowyrazowymi są np. *chilli con carne* jako nazwa potrawy czy *garam masala* jako nazwa przyprawy.

Jeśli w Słowosieci już jest coś, co nie powinno tam być, bo jest niezgodne z powyższą regułą, to to tam zostawiamy. Takie obce jednostki mogą mieć status: *Sprawdzony*. Nie wprowadzamy jednak nowych jednostek obcych, o obcej pisowni, niezadomowionych w polskim systemie językowym, a jeśli wprowadzamy coś, co uznamy za zadomowione, warto poprzeć naszą decyzję krótkim komentarzem (np. „Częste w KGR10").

Imiesłowy

Imiesłowy traktujemy jako formy czasownika i nie wprowadzamy ich jako osobnych lematów do Słowosieci. W poprzednich wersjach Słowosieci były one dopuszczane i łączone z czasownikiem relacją synonimii międzyparadygmatycznej V-Adj. Relacja ta już nie istnieje, a pozostałości po starym systemie, tzn. jednostki imiesłowowe, usuwamy.

Kwestia odróżniania imiesłowów od przymiotników została opisana w instrukcji przymiotnikowej. Zgodnie z tymi wytycznymi imiesłów przede wszystkim musi mieć końcówkę imiesłowową (*-ny, -ty, -ący*) i wiązać się z wyrażeniami temporalnymi (np. *teraz*, *w tej chwili*, *wczoraj* czy *na długo*). Zgodnie z tymi wytycznymi[1] formy z końcówką *-ły* traktujemy jako formy zadiektywizowane (jest to pozostałość po imiesłowie historycznym, Słowosieć dąży do opisu synchronicznego).

Jednak kwestia imiesłowów okazuje się nietrywialna i wciąż pojawiają się w Słowosieci formy, które mogą być uznawane za imiesłowy. Formy te traktujemy jako błędne, tzn. przyznajemy im status: *Błąd* i komentujemy odpowiednio (np. „Prawdopodobny imiesłów"). Imiesłowy będą usuwane.

Pewną pomocą w ocenie, czy coś jest, czy nie jest imiesłowem, może być też tager (https://ws.clarin-pl.eu/tager.shtml), jednak należy pamiętać, że funkcją tagerów jest analiza kontekstowa (tzn. ujednoznacznianie morfologiczne może różnić się w zależności od kontekstu, w którym występuje interesujące nas słowo). Tager MorphoDiTa pracuje w modelach dla języka współczesnego i historycznego (XIX-wiecznego).

[1] Zob. też. B. Bartnicka (1970): Adiektywizacja imiesłowów w języku polskim, Warszawa: PWN.

Gerundia

Gerundia traktujemy jak formy czasownikowe. Zostały one wprowadzone do Słowosieci 2.0 automatycznie w ramach eksperymentu, a następnie były weryfikowane przez lingwistów.

Cechą charakterystyczną tych form jest synonimia międzyparadygmatyczna V-N oraz miejsce w drzewie gerundialnym, które ma w najwyższej partii synset sztuczny GERUNDIUM oraz, niżej, synsety sztuczne, które są wierną kopią sztucznych synsetów czasownikowych ze Słowosieci 2.0. Dziedzinami gerundiów są [zdarz] dla gerundiów tworzonych od czasowników dokonanych, [czy] dla gerudniów od czasowników niedokonanych oraz [st] dla gerundiów od stanów. Najczęściej mają również w komentarzu synsetu AOds (*automatyczny odsłownik*) i autora GerGer, jednakże komentarz ten mógł się zmienić w trakcie pracy na synsetach gerundialnych w ramach opisanego eksperymentu, dlatego nie jest on wyznacznikiem. Najłatwiej gerundium rozpoznać po synonimii międzyparadygmatycznej.

Od czasu tego eksperymentu na gerundiach nie były prowadzone żadne systematyczne operacje (poza indywidualną inicjatywą niektórych Lingwistów). W tej chwili opis gerundiów nie jest w żaden sposób wspierany, formy te nie wchodzą też do żadnych statystyk w Słowosieci i nie powinny istnieć poza swoim drzewem. Niewykluczone, że gerundia będą kiedyś aktualizowane automatycznie tak, by były kompatybilne z drzewami czasownikowymi, jednak z obecnej perspektywy wydaje się to nieużyteczne. W związku z tym nie poprawiamy gerundiów, nie dajemy im nowych komentarzy, nie błędujemy. Wyjątek stanowią gerundia w drzewach rzeczownikowych, które powinny otrzymywać status: *Błąd* i być przenoszone do drzew gerundialnych.

Wykrzykniki

Wykrzykniki nie należą do części mowy opisywanych w Słowosieci (wykrzyknik nie jest rzeczownikiem) i jako takie powinny być systematycznie usuwane. Nie wprowadzamy nowych wykrzykników do Słowosieci.

Również w tym wypadku, jeśli istnieją wątpliwości co do oceny, czy coś jest, czy nie jest wykrzyknikiem, pomocą może służyć tager (https://ws.clarin-pl.eu/tager.shtml), które powinny być oznaczane jako *interj*.

Liczba mnoga i pojedyncza

Dla większości lematów formą podstawową jest forma w liczbie pojedynczej. Wyjątek stanowią *pluralia tantum* - w tym przypadku to, oczywiście, forma w liczbie mnogiej jest tą podstawową, którą wprowadzamy do Słowosieci. Jeśli więc istnieje forma w liczbie pojedynczej, powinniśmy ją wprowadzić i opisać, formę w liczbie mnogiej traktując jako fleksyjną. Jednostką leksykalną nie będą więc np. *plusy* w znaczeniu szkieł korekcyjnych (jest *plus* - jedno szkło; przykład z komentarza jednostki: *Na jednym oku ma plus, na drugim minus*), ale będą *okulary,* bo *okular* i *okulary* to jednostki o innych znaczeniach. Jednostką nie mogą być w takiej interpretacji również *pierogi ruskie*, ponieważ to danie to kolekcja pojedynczych bytów o nazwie *pieróg ruski* (jest w Słowosieci). W przypadku wielowyrazowych jednostek leksykalnych przed usunięciem ich

z bazy lub wprowadzeniem do niej nowych należy wziąć pod uwagę szereg innych wyznaczników, opisanych w akapicie Wielowyrazowe jednostki leksykalne.

Liczba mnoga preferowana jest w przypadkach, kiedy możemy sobie wyobrazić liczbę pojedynczą, jednak w praktyce językowej (w tekstach) jest ona bardzo rzadka, wydaje się nienaturalna lub wręcz - nie istnieje. Jako przykład można podać obuwie takie jak *brogsy*, *derby* czy *wiedenki*. Może również zaistnieć sytuacja, w której liczba pojedyncza zacznie być używana. Wtedy powinno się rejestrować jedynie formę podstawową. Słowosieć to twór dynamiczny, a każda kolejna wersja ma w założeniu jak najwierniej oddawać aktualny stan języka. Jest więc zupełnie naturalnym, że niektóre formy będą zastępowane innymi.

Nie wprowadzamy do Słowosieci

Poza kategoriami opisanymi powyżej, nie wprowadzamy do Słowosieci:

1) imion i przymiotników dzierżawczych od nich,
2) nazw własnych niebędących podstawami derywacyjnymi,
3) innych części mowy, poza rzeczownikiem, czasownikiem, przymiotnikiem i przysłówkiem.

Nieopracowane jednostki

Komplet znaczeń

Jeśli ktoś z Lingwistów czuje potrzebę opracowania danego lematu, to należy pamiętać, że przy weryfikacji pracujemy z kompletem znaczeń, tzn. nie poprawiamy komentarza dla jednej jednostki. Jest to dopuszczalne w wyjątkowych sytuacjach, kiedy jednostka jest potrzebna do opracowania czegoś innego, wchodzi w skład siatki relacji dla jakiegoś innego znaczenia. Wtedy należy pozostawić status: *Znaczenie*, jednak staramy się, żeby było jak najmniej *Znaczeń* w Słowosieci, pracujemy raczej z całymi strukturami, niż pojedynczymi jednostkami.

Kompletowi znaczeń edytujemy komentarze wg zasad opisanych w dokumencie Komentarze w Słowosieci, a także relacje synsetów i jednostek, sprawdzając jednocześnie skład synsetów. Zasady pracy z kompletem znaczeń zostały opisane w wytycznych dla czasownika (rozdz. VI - Procedura).

Stary format komentarza

Stary format komentarza nie jest błędem i nie należy go tak traktować w systemie korekty Słowosieci. Świadczy on najczęściej o tym, że dana jednostka, a także komplet znaczeń dla danego lematu, nie były weryfikowane od momentu wprowadzenia ich do Słowosieci przez Lingwistę. Mogą mieć błędy (ale nie muszą).

Słowosieć jest weryfikowana i uzupełniana podczas różnych projektów, najczęściej nastawionych na konkretne zastosowania lub konkretny obszar badawczy i nie sposób wszystkiego zrobić od razu. Najczęściej są to zmiany systematyczne. Jeśli ktoś z Lingwistów potrzebuje

zweryfikować i poprawić komplet znaczeń dla lematu, którego jednostki mają stary format komentarza, to należy to zrobić zgodnie z powyższymi zasadami.

Relacje

Przechodniość relacji

Relacjami, które zapewniają w Słowosieci przechodniość relacji, są hipo- i hiperonimia, tzn. dany synset dziedziczy wszystkie składniki znaczenia (*genus proximum*) po swoim hiperonimie. Jednak tylko hipo- i hiperonimia są przechodnie w całych liniach tej relacji, tzn. najniższy hiponim w drzewie powinien być hiponimem nie tylko swojego bezpośredniego hiperonimu, ale także tego najwyższego.

Oprócz tego, jak zostało powiedziane powyżej, istnieją takie relacje, które wystarczają do opisu jednostki lub synsetu dzięki temu, że jednostka lub synset „dziedziczą" relacje po jednostce lub synsecie, do której relacja prowadzi (zob. Relacje wymagane).

Stare relacje

W przypadku relacji, które już nie istnieją oraz takich, które zmieniły swoją definicję, niezbędna jest kompleksowa ich poprawa. Do takich relacji zaliczyć można, między innymi, aspektowość wtórną, meronimię i holonimię, mero- i holonimię czasownikową. Poprawa tych relacji będzie się odbywać na zasadzie zadań zorientowanych wokół danej relacji. Dopuszczalne (i wskazane) jest poprawianie instancji tych relacji na bieżąco. Synset ani jednostka nie mogą otrzymać statusu *Sprawdzony*, jeśli relacje, o których mowa, nie zostały poprawione.

Stan\|cecha a Cecha definicyjna

Cecha definicyjna została wprowadzona w nowych wytycznych dla rzeczownika i miała odpowiadać relacji stanu|cechy, jednak na poziomie synsetów. Stare relacje Stanu|cechy miały zostać automatycznie zastąpione nowymi relacjami Cechy definicyjnej. Praktyka pokazała jednak, że ta decyzja nie była najtrafniejsza - Słowosieć jest dobrze skonstruowanym organizmem, w którym informacje są powiązane w sieć i żadna z nich nie jest zbędna ani nieprzydatna. Potwierdziły to analizy z artykułu Dynamic verbs in the Wordnet of Polish, który na przykładach pokazuje logiczną zależność niektórych relacji w Słowosieci. Z perspektywy informatywności relacji nie ma przesłanek za tym, żeby uważać te dwie relacje za tautologiczne. Z tego powodu można rozważyć wyniesienie innych relacji derywacyjnych na poziom synsetów (co częściowo zaczęliśmy robić dla czasownika i przymiotnika) lub, odwrotnie, wskazać na możliwości powiązań derywacyjnych jednostek z synsetów, które są ze sobą powiązane relacjami.

Istnienie relacji Stanu|cechy w Słowosieci nie należy więc uznawać za błędne.

Antonimia właściwa

Uznajemy, że wszystkie jednostki w danym synsecie powinny mieć tę samą antonimię. Jeśli dwie jednostki mają takie same znaczenie, to, na poziomie semantycznym, powinny wchodzić w takie same związki ze słowami, które mają znaczenie przeciwne.

Jednak w niektórych przypadkach antonimia występuje tylko pomiędzy jednostkami i nie jest rozszerzana na ich synonimy. Są to:

1) antonimy kulturowe (np. *anioł* - *diabeł*), które jednocześnie w różnych opracowaniach nt. Słowosieci są jednym z głównych argumentów za tym, żeby antonimia była na poziomie jednostek, nie synsetów;
2) słowa prefiksowane, w których antonimia wiąże się z semantyką prefiksu, np. *tlenowy* i *beztlenowy*, *areobowy* i *anaerobowy*.

Klasy i dziedziny

Do tej pory nie zostały stworzone wytyczne dla dziedzin semantycznych w Słowosieci. Ich przydzielanie jest intuicyjne i zależy od lingwisty. Dziedziny semantyczne powinny jak najtrafniej definiować znaczenia, ale nie jest do tej pory określone, jak oceniać tę „trafność". Jedynymi przypadkami, w których dziedziny zostały skodyfikowane, są trzy dziedziny przymiotnika i klasa stanowa czasowników (dziedzina [cst]). Na potrzeby opisu czasownika stworzono dodatkowe dziedziny dla klas pomocniczych: [cdystr], [caku], [cdel] i [cper]), jednak badania wykazały, że klasy pomocnicze nie mają dużego znaczenia przy opisie czasowników, zwłaszcza jeśli chodzi o rozróżnianie znaczeń polisemicznych.

Kwalifikatory

W dokumencie Kwalifikatorów system zostało wskazane 11 kwalifikatorów, stosowanych w Słowosieci na oznaczenie rejestru: daw. (dawny), reg. (regionalny), specj. (specjalistyczny), środ. (środowiskowy), książk. (książkowy), urz. (urzędowy), wulg. (wulgarny), posp. (pospolity), pot. (potoczny), og. (ogólny) oraz nienorm. (nienormatywny). Kwalifikatory te wyczerpują zakres rejestrów opisywanych w Słowosieci. Wskazany dokument określa ponadto zakres ich stosowalności.

W przypadku słów nienormatywnych, dawnych oraz regionalizmów problem kwalifikatorów jest zbieżny z problemem słów, których nie opisujemy. Słowosieć jest narzędziem opisu synchronicznego, dlatego też słownictwo dawne nie występuje regularnie - jest wprowadzane w przypadku wysokiej frekwencji w korpusach lub na użytek konkretnych zastosowań. Nie wprowadzamy archaizmów. Wspomniany powyżej dokument definiuje wyraz dawny jako taki, który niedawno wyszedł lub wychodzi z użycia. Wyrazy starsze, których znaczenie da się zrekonstruować jedynie na podstawie ich obecności w starych słownikach, niemające potwierdzenia w używanych XIX, XX lub XXI wieku tekstach, są uznawane przez nas za archaizmy.

Podobnie nie opisujemy regularnie słownictwa regionalnego - interesują nas jedynie te jednostki leksykalne, które mają wysoką frekwencję (powyżej 50 wystąpień) w korpusach. W

przypadku regionalizmów warto zaznaczyć w glosie, z jakiego regionu dane znaczenie pochodzi, jeśli taka informacja jest możliwa do dodania na podstawie ogólnodostępnych tekstów (bez specjalistycznych kwerend).

Za pomocą kwalifikatora *nienorm.* oznaczamy jednostki leksykalne, które nie zostały uznane za zgodne z normą językową przez autorytety (Rada Języka Polskiego, słowniki poprawnościowe, poradnie językowe), jednak istnieją w polskim systemie językowym i są używane w mowie potocznej. Nie wprowadzamy do Słowosieci oczywistych omyłek językowych, literówek, słów bez znaków diakrytycznych czy z błędami ortograficznymi.

Jednostka nienormatywna, zanim zostanie wprowadzona do Słowosieci, musi zostać potwierdzona przez korpus (w tym przypadku wprowadzamy jednostki, które mają liczbę wystąpień powyżej 50 w analizowanym korpusie). Nie wprowadzamy jednostek nienormatywnych spoza korpusu.

Słowa potoczne, pospolite i wulgarne

W dokumencie Kwalifikatorów system zostały wskazane kryteria rozróżnienia tych rejestrów, jednak, mimo to w praktyce to rozróżnienie sprawia problemy. Zgodnie z przyjętymi zasadami potocyzmy mogą być używane w sytuacji publicznej i nie są traktowane przez odbiorcę jako niedostosowane do tej sytuacji, podczas gdy jednostki z pozostałych dwóch rejestrów mają wyraźne nacechowanie pragmatyczne: świadczą o skróceniu dystansu przez nadawcę, nieumiejętności dostosowania słownictwa. Aby pomóc lingwistom w rozróżnieniu słów należących do tych rejestrów, został przywołany eksperyment myślowy związany z sytuacją, w której spotykamy w miejscu publicznym nieznajomego. Chociaż w praktyce zastosowanie danego słowa jest indywidualną kwestią rozmówcy, Słowosieć powinna opisywać najbardziej prototypowe i intersubiektywne użycie, nie zaś indywidualne preferencje lingwisty lub jego otoczenia. W związku z tym uznajemy, że słowa nacechowane, których możemy użyć w sytuacji publicznej, w rozmowie z nieznajomym, którego statusu nie znamy, otrzymają kwalifikator *pot*. Słowa, które będą odbierane w takiej sytuacji jako skrócenie dystansu, niedostosowanie słownictwa, będą kwalifikowane jako *posp.* lub *wulg.*, przy czym tych pierwszych możemy użyć w sytuacji familiarnej, np. w gronie rodziny. Słowa kwalifikowane jako *wulg.* są jednoznacznie oceniane jako wulgarne. W związku z tym do jednej lub drugiej grupy będzie należała znaczna część słów obraźliwych (np. *debil*) używanych w języku polskim. Nie jest dopuszczalne, aby Słowosieć opisywała tego typu słowa jako możliwe do użycia w sytuacji publicznej, w kontakcie z nieznajomym.

Eufemizmy i słowa przenośne

Przed wprowadzeniem systemu kwalifikatorów do Słowosieci opisywano w niej użycia eufemiczne, stosując kwalifikator *euf.* lub *eufem.* Obecnie nie ma możliwości ich użycia z powodu niezgodności z przyjętym systemem. Słowa ocenione jako eufemizmy powinny być poddane testowi na przynależność do rejestru zgodnie z dokumentem Kwalifikatorów system.

Podobnie we wcześniejszych wersjach używany był kwalifikator *przen.*, który oznaczał znaczenie przenośne. Nie ma go w obecnym systemie, a skrótu tego nie powinno się używać w takiej formie ze względu na zbieżność z innymi słownikami, w których słowa przenośne są wyodrębniane. Obecnie jednostki zawierające w komentarzu skrót *przen.* należy poddać testom na przynależność do rejestru. Bardzo często będą miały one w nowym systemie kwalifikator *książk.*, choć nie jest to regułą.

Słowa środowiskowe a specjalistyczne

We wspomnianych wytycznych do anotacji rejestrem zostały również wskazane różnice pomiędzy rejestrami specjalistycznym i środowiskowym. W zakres słownictwa specjalistycznego wchodzi nie tylko terminologia naukowa typowa dla różnych dziedzin nauki, ale także słowa używane przez fachowców (np. rybaków, leśników, rolników, hydraulików) oraz hobbystów (np. graczy komputerowych, miłośników sztuki, jeśli nie tworzą oni subkultury), czyli te zakresy słownictwa, które można poznać uczestnicząc w danym dyskursie i spotykając się z innymi ich uczestnikami. Wyjątek stanowią zamknięte grupy społeczne, do których trudno się dostać (oprócz wymienionych w instrukcji więźniów będą to np. hackerzy, czy działające legalnie tajne stowarzyszenia) oraz subkultury (np. *skejci*, którzy, oprócz tego, że są hobbystami, tworzą subkulturę), których słownictwo będzie kwalifikowane jako środowiskowe. Przyjmujemy, że hobbyści różnią się od uczestników subkultury tym, że subkulturę da się nazwać i zdefiniować ramy jej dyskursu.

Jednostki anotowane na potrzeby WSD

Jednostki anotowane w korpusach na potrzeby ujednoznaczniania znaczeń (WSD, Word Sense Disambiguation) mają oznaczenia WSD w komentarzach synsetów. Anotacja WSD jest zawsze prowadzona przy pomocy określonej wersji Słowosieci.

Słowosieć ewoluuje, język też. Z tej perspektywy trzymanie „na stałe” jednostek WSD w niezmiennym znaczeniu i z niezmiennymi relacjami w odniesieniu do stanu, w którym dana jednostka była wykorzystana do anotacji, wydaje się niepraktyczne i przeczące zasadom tej ewolucji. Ponadto prowadzone są rejestry zmian znaczenia jednostek WSD w kolejnych wersjach Słowowieci, a korpusy, wykorzystujące Słowosieć do anotacji (oprócz KPWr jest to też Składnica) okresowo aktualizowane względem nowej wersji.

Nie istnieje więc żaden powód, dla którego nie należy dokonywać zmian w jednostkach oznaczonych jako WSD.

Synsety sztuczne

W Słowosieci istnieje system synsetów sztucznych, czyli takich, które nie są jednostkami języka naturalnego. Organizują one strukturę hierarchiczną Słowosieci i są dodawane przez Lingwistów w razie potrzeby. Przyjmujemy, że żeby móc dodać sysnet sztuczny, musi on mieć wystarczającą liczbę hiponimów, przy czym to, co „wystarczające” nie jest nigdzie definiowane.

Niektóre sysnety sztuczne powstały w ten sposób, że stały się nimi połączenia wielowyrazowe, uznane za nie-jednostki, które miały już hiponimy i były zakotwiczone w strukturze Słowosieci za pomocą innych relacji. Synsety sztuczne nie są wliczane do statystyk Słowosieci.

Synsety sztuczne były regularnie dodawane w przypadku czasownika, który opisywano w Słowosieci 2.0 za pomocą klas Romana Laskowskiego i Zeno Vendlera. Synsety te odwzorowywały te klasy. W wersji 4.0 Słowosieci zrezygnowaliśmy z klas Vendlera-Laskowskiego – synsety przestały pełnić swoją funkcję (stare klasy pojawiają się w komentarzu jako VLC) i można je usuwać. Jednak należy zachować szczególną ostrożność przy tej pracy. Zalecane jest oznaczanie synsetów sztucznych z klasami czasowników za pomocą statusu *Błąd* i komentarza „Do usunięcia".

Wielowyrazowe jednostki leksykalne

Derywacyjność od jednostek wielowyrazowych

Jednostki wielowyrazowe mogą pełnić rolę podstaw słowotwórczych na równych prawach, co leksemy jednowyrazowe. Jako przykład bardziej oczywistych można podać przymiotniki derywowane od wieloczłonowych nazw własnych: północnoamerykański - Ameryka Północna, południowokoreański Korea Południowa itp., czy czasowniki z „się" będące podstawami derywacyjnymi, np. *ześlizgiwać się - ześlizg*. Inne często występujące przykłady to derywowanie przymiotnika od dwuczłonowego terminu, np. *geografia afiniczna - afiniczny*, czy derywowanie nazwy pospolitej od nazwy własnej, np. *Adam Słodowy* jako człowiek sprawny technicznie, który potrafi wykonać przydatne konstrukcje i narzędzia z rzeczy codziennego użytku derywowany od nazwiska Adama Słodowego, bohatera popularnego programu telewizyjnego.

Derywaty odfrazeologiczne są opisane na przykład w artykule Od frazeologizmu do derywatu. Zgodnie z opisaną powyżej tendencją powinniśmy również jako derywaty odfrazeologiczne traktować słowa typu *wodolejstwo* (od *lać wodę*), czy *dziwowisko* (od *dziwować się*).

Relacja derywacyjności nie określa, z jaką derywacją mamy do czynienia, więc jako derywaty w Słowosieci są klasyfikowane zarówno wytwory derywacji semantycznej, za pomocą której tworzy się polisemy (chociażby wspomniany *Adam Słodowy*, czy tworzenie ilości, zawartości od nazwy substancji np. *mąka* jako produkt spożywczy i *mąka* jako torebka tego produktu) oraz słowa różniące się na poziomie morfologicznym od swych podstaw (np. przymiotniki odrzeczownikowe, dla których nie ma bardziej szczegółowej relacji w wytycznych). Przy dodawaniu relacji derywacyjności należy jednak pamiętać, że Słowosieć zawiera opis synchroniczny języka, tak więc jedyną relacją dla wskazania zależności diachronicznych może być fuzzynimia.

Pojęcia specjalistyczne i działy nauk

„Meta" definicje

W Słowosieci istnieją pojęcia specjalistyczne, mające niską frekwencję w języku ogólnym, często abstrakcyjne i trudne do zdefiniowania. Trend rozwoju Słowosieci, motywowany potrzebami jej użytkowników, wskazuje, że tego typu słownictwo będzie coraz częściej opisywane. Jest to drugi, po języku potocznym, np. z social mediów, perspektywiczny kierunek rozwoju. Słownictwo specjalistyczne umożliwia konstruowanie narzędzi do analizy tekstów specjalistycznych na różnym poziomie przetwarzania, w tym także - opartych na ontologiach dziedzinowych, z którymi Słowosieć jest i będzie łączona w ramach różnych projektów.

Jednostki leksykalne należące do rejestru specjalistycznego powinny być jednak opisywane jako elementy rzeczywistości - wykluczamy używania w glosach składników „meta" opisu. Do takich należy *pojęcie* jako *genus proximum* w przypadkach, kiedy dana jednostka nie jest pojęciem, np. *macierz 1*. Słowo pojęcie można w takich definicjach zastąpić formułą opisową, np. „*w matematyce: …*" Podobną funkcję metaopisową pełnią relacje hiponimii do synsetu *pojęcie 2|kategoria 3* oraz ich sztucznych hiponimów.

W nowej instrukcji rzeczownikowej zostały zdefiniowane trzy relacje, które mogą okazać się szczególnie przydatne w definiowaniu tego typu jednostek: *cecha definicyjna*, *obszar* oraz *parametr*. Relacje z hiponimii z pojęciem należy zamieniać na fuzzynimię w przypadku, kiedy definiowana jednostka nie jest pojęciem i kiedy nie można tej relacji zamienić na inną.

Znaczenie jednostki specjalistycznej a zakres stosowalności

Dyscyplina naukowa, w której się stosuje jakiś termin, określająca zakres jego stosowalności, nie może być jedynym wyznacznikiem definiowania znaczenia i tylko na jej podstawie nie powinno się wnioskować o występowaniu osobnego znaczenia. Szczególnie należy zwrócić uwagę na przypadki, kiedy na gruncie dziedziny dokonuje się jedynie definiowania właściwości jakiegoś bytu, którego nazwę stosuje się również w języku ogólnym lub kiedy w różnych dziedzinach określa się w różny sposób właściwości tego samego bytu. Na przykład *romantyzm 3* jest definiowany jako *kierunek w sztuce* i błędem byłoby wyodrębniać osobny *romantyzm* w literaturze, muzyce, malarstwie itp. Z kolei *ewolucjonizm* w biologii jako element teorii ewolucji i *ewolucjonizm* w naukach społecznych i politycznych, próbujący znaleźć w rozwoju społeczeństw regularności podobne do rozwoju biologicznego gatunków, to terminy naukowe oddające dwie różne koncepcje.

Nazwy własne i wyrazy pospolite

Nazwy własne w Słowosieci

W Słowosieci nazwy własne nie są regularnie opisywane. Wyjątek stanowią takie, od których są derywowane jednostki pospolite. Wszystkie nazwy własne mają w polu komentarza obowiązkowy znacznik NP (na początku komentarza), mogą również mieć glosę i link do

Wikipedii. Kwalifikator dodaje się jedynie w przypadkach, kiedy nazwa własna jest nacechowana stylistycznie (np. potoczna *Wawa*).

Relacją powołaną do opisu hierarchii synsetów zawierających nazwy własne jest relacja typu, łącząca nazwę ze słowem pospolitym, którego nazwa jest egzemplarzem (i obligatoryjnie odwrotna: egzemplarza). Mimo to niektóre nazwy własne mogą mieć hiponimię lub hiperonimię. Dzieje się tak w przypadkach, kiedy relacja łączy dwie nazwy własne o stosunku nadrzędności/podrzędności. Poza tym wyjątkiem nazwy własne mogą wchodzić z innymi nazwami własnymi w inne relacje opisane w instrukcjach.

Problemy szczegółowe nazw własnych:

1) nazwy plemion w liczbie mnogiej nie są nazwami własnymi; słowa typu: Irokezi, Masajowie, Polanie, to liczba mnoga od nazw etnicznych (które również nie są nazwami własnymi) i powinny być usuwane zgodnie z zasadami stosowania liczby mnogiej i pojedynczej;
2) nazwy okresów geologicznych również nie są nazwami własnymi, to deskrypcje określne;
3) jeśli występują dwa warianty ortograficzne zapisu – z małej i dużej litery – włączamy oba warianty jako oddzielne jednostki.

Problemy szczegółowe poszczególnych części mowy

Czasowniki

Czasowniki statyczne

Przygotowując procedurę opisu relacjami czasowników w Słowosieci 4.0 założyliśmy, że różnice pomiędzy klasą czasowników stanowych a klasą czasowników dynamicznych będą na tyle wyraźne, że dadzą się ująć za pomocą ram opisu typowego dla Słowosieci. Bardziej szczegółowe badania wykazały jednak, że są takie obszary semantyki czasownika, w których nie można jednoznacznie wnioskować o klasie. Badania wykazały także, że możemy mówić o znaczeniach prototypowo stanowych i prototypowo dynamicznych. Niektóre ze znaczeń nie dają się jednak jednoznacznie zaklasyfikować do żadnej z tych dwóch klas. W ich przypadku o tym, czy czasownik jest dynamiczny, czy statyczny świadczą konteksty użycia, można więc mówić o znaczeniu kontekstowym, a nie leksykalnym.

Podział na klasy czasowników jest ważny ze względu na relacje. Zakładaliśmy, że pomiędzy czasownikiem stanowym a dynamicznym nie może być aspektowości czystej, ponieważ ta wskazuje na to, że znaczenie obu czasownikach w parze aspektowej jest to samo. W tych przypadkach, w których ciężko jest określić, czy czasownik jest statyczny, czy dynamiczny, dopuszczamy aspektowość czystą pomiędzy parą aspektową, o ile zachowana jest zgodność z testami dla tej relacji. Przykładem takiego czasownika jest *oblewać 8*.

Zasadą Słowosieci było również to, że czasowniki powinny się łączyć w drzewa hiperonimiczne jedynie w obrębie klasy. W przypadku czasowników, których stanowość lub dynamiczność wynika z kontekstu, a nie jest cechą leksykalną, dopuszczamy również hiponimię

pomiędzy statycznym i dynamicznym, o ile testy podstawieniowe zostaną spełnione. Jednak to odstępstwo dotyczy jedynie tych czasowników wątpliwych, co do których ciężko orzec, czy są statyczne, czy dynamiczne. W przypadku pozostałych czasowników stosujemy zasadę łączenia w drzewa w obrębie klas.

Nadal utrzymujemy zasadę, że czasowniki prototypowo stanowe, tj. takie, dla których stanowość jest cechą leksykalną, nie kontekstową, powinny mieć dziedzinę [cst] i nie mogą mieć relacji typowych dla czasowników dynamicznych, tj. kauzacji, procesywności oraz inchoatywności. Podobnie czasowniki prototypowo dynamiczne nie mogą mieć relacji stanowości.

Aspektowość a rejestry

Bywają przypadki, w których czasownik w jednym z aspektów funkcjonuje w języku ogólnym, a w drugim ma ograniczony zakres stosowalności (np. zachował się tylko w dialektach lub w tekstach dawnych). Czasowniki, które są dla siebie parą aspektową, ale mają nielubiące się kwalifikatory, mogą być połączone aspektowością czystą, ponieważ nie interpretujemy rejestru jako różnicy semantycznej, tylko informację pragmatyczną.

Rzeczowniki

Feminatywy i żeńskie nazwy własne

Relacja żeńskości jest wystarczająca do opisu feminatywów i żeńskich nazw własnych, jeśli jednak możliwe jest stworzenie drzewa żeńskiego, dopuszczalne jest stworzenie go. W systemach NLP hiponimia ma większą wartość, niż relacja żeńskości.

Nazwy mieszkańców

Nie każdy mieszkaniec danego miasta ma przynależność narodową do kraju, w którym leży to miasto, więc nie powinno być hiponimii pomiędzy mieszkańcami, a narodowościami. Dla rzeczowników określających mieszkańców został powołany synset sztuczny „mieszkaniec miasta”.

Nacechowanie

Jednostki leksykalne, które są derywatami, łączonymi z podstawami za pomocą relacji nacechowania, nie mogą być ze swoimi podstawami słowotwórczymi w synsetach, ponieważ z definicji nacechowania wynika, że relacja ta pociąga za sobą zmianę znaczenia w stosunku do podstawy. W przypadku takich par derywacyjnych, pomiędzy którymi nie zachodzi zmiana znaczenia, stosujemy derywacyjność.

Przymiotniki i przysłówki

Stopień wyższy i najwyższy

Przymiotniki i przysłówki w stopniu wyższym i najwyższym, które są tworzone regularnie, nie powinny być wprowadzane do Słowosieci. W jej wcześniejszych wersjach były one wprowadzane dla częstych znaczeń przymiotnikowych i łączone relacjami Stopień wyższy i Stopień najwyższy, które pozostały w obecnej wersji. Pozostały też niektóre z tych jednostek przymiotnikowych - są one w tej chwili wykorzystywane w systemach WSD oraz do znakowania korpusów i nie powinny być usuwane. Jeśli w przymiotniku w stopniu wyższym lub najwyższym pojawia się błąd (związany np. z kompletnością znaczeń), należy go poprawić. Nie wprowadzamy jednak nowych przymiotników w stopniu wyższym i najwyższym.

-ości

Tzw. *-ości* to odprzymiotnikowe rzeczowniki transpozycyjne, tworzone od podstaw słowotwórczych za pomocą końcówki *-ość* i łączone z nimi za pomocą synonimii międzyparadygamtycznej. Rzeczownik transpozycyjny odprzymiotnikowy musi mieć dokładnie takie same cechy semantyczne, jak przymiotnik - różni się tylko cechami gramatycznymi. W związku z tym musi dać się scharakteryzować za pomocą parafrazy „X-owość cecha kogoś lub czegoś, kto jest Y-owy lub co jest Y-owe". Rzeczownik taki będzie wchodził w kolokacje z tymi samymi rzeczownikami, co przymiotnik, od którego jest derywowany (np. piękny głos = *piękność głosu*). *-Ości* zawsze mają dziedzinę [cech].

Wykluczyliśmy możliwość tworzenia rzeczowników transpozycyjnych odrzeczownikowych za pomocą końcówek innych niż *-ość*, ale nie wykluczamy, że mogą one znajdować się w synsecie z nie-*ościami*. Podstawą łączenia w synsety są relacje synsetów, nie jednostek, wyrażające tożsamość znaczeniową. Nie możemy więc arbitralnie wyłączyć *-ości* z synsetu.

Nie jest błędem, jeśli jakiś przymiotnik nie posiada swojej *-ości*. W odróżnieniu od gerundiów, *-ości* nie były tworzone automatycznie i są pełnoprawnymi jednostkami Słowosieci, liczonymi w statystykach i wprowadzanymi na podstawie wystąpień w korpusie.

Kolokacyjność i kolokacje

Kolokacyjność jest relacją jednostek, która została powołana do życia na potrzeby projektu NCBR. Oddaje ona związki syntagmatyczne, czyli tendencję do wchodzenia jednostek leksykalnych w kolokacje rozumiane jako częste połączenia wyrazowe w korpusie tekstów. Dzięki relacji te połączenia można ujdenoznacznić na poziomie jednostkowym. Oprócz relacji w tym czasie powstało w WordNet Loomie pole do wpisywania kolokacji w atrybuty jednostki, czym zrezygnowano ze stosowania kolokacji zamiast przykładów użycia w polu komentarza jednostki.

W dokumencie: Procedura dodawania kolokacji opisano warunki, jakie powinna spełniać kolokacja, jednak, jak już zostało powiedziane, był on tworzony na potrzeby biznesowego projektu. Obecnie w stosunku do kolokacji panuje nadrzędna zasada, że ma być jednoznaczna i jak najwierniej oddawać sens danej jednostki. Nie musi być neutralna pod względem wydźwięku. Dla niektórych jednostek - jest to wręcz niewskazane. Poza tym idealnie jest, gdy kolokacja da się opisać za pomocą relacji kolokacyjności (z punktu widzenia WSD relacja jest dużo ważniejsza,

niż kolokacja zapisana w formie tekstowej), więc nie może być zbyt złożona. Należy również pamiętać, że kolokacje to nie to samo, co frazy.

We wcześniejszych wersjach Słowosieci kolokacje mogły być stosowane zamiast przykładów użycia. W niektórych przypadkach, np. przymiotników relacyjnych, czy też jednostek o ograniczonej łączliwości, dobrze dobrana kolokacja może zilustrować znaczenie równie trafnie, co przykład użycia. W takich wypadkach dopuszczamy dodawanie kolokacji i relacji kolokacyjności zamiast przykładów użycia.

Błędy anotacji i rzutowania

Rzutowanie na Princeton WordNet

Słowosieć jest rzutowana na Princeton WordNet za pomocą relacji międzyjęzykowych od 2012 roku. Od rozpoczęcia tego procesu zostały zrzutowane prawie wszystkie rzeczowniki, większość przymiotników i przysłówków. Praca nad czasownikami rozpoczęła się w 2018 roku i w tej chwili trwa. Rzutowanie jest prowadzone głównie na poziomie synsetów, ale w 2018 i 2019 roku prowadzono również pilotażowe prace łączenia obu wordnetów na poziomie jednostek różnymi typami ekwiwalencji międzyjęzykowej. Tych drugich jest niewiele.

W związku z tym, że prace zarówno zespołu anglistów (dodawanie nowych relacji międzyjęzykowych), jak i polonistów (rozszerzanie Słowosieci) prowadzone są w trybie ciągłym, mogą istnieć takie synsety, które są niezrzutowane. To, że jakiś synset nie posiada relacji międzyjęzykowej, nie jest błędem. Podobnie jak błędem nie jest jednostka bez takiej relacji. Żadne z nich nie powinno być znakowane jako *Błąd*. Ponadto jednostki i synsety bez relacji międzyjęzykowych mogą otrzymać status: *Sprawdzony*.

Podobnie jak po polskiej stronie Słowosieci, w rzutowaniu zdarzają się błędy. Można je odnotowywać w tej tabeli: https://docs.google.com/spreadsheets/d/1WwAaDdwK25umCDSQh-zdCJjSzXKF_tzil8FY9ef_Qjo/edit?usp=sharing lub znakując *Błąd* w komentarzu dla całego angielskiego synsetu. Dzięki temu drugiemu sposobowi odnajdywanie błędów może odbywać się w sposób bardziej zorganizowany, ale tabele umożliwiają szybszą reakcję. Nie ma potrzeby korzystania i z tabeli, i z komentarza do synsetu angielskiego.

Należy uważać na potencjalne odpowiedniki - są to automatyczne relacje pomiędzy wordnetami polskim i angielskim, oznaczane na grafie jako „po_pa” i „po_ap”, używane jako podpowiedzi dla lingwistów rzutujących. Nie są one faktycznymi synonimami międzyjęzykowymi, dodanymi ręcznie, i są usuwane po każdej iteracji rzutowania, w której są wykorzystywane. Mogą w nich być błędy, ale nie oznaczamy takich synsetów statusem *Błąd*.

O ile możemy poprawić błędy po stronie polskiej i w rzutowaniu, o tyle - nie jesteśmy w stanie poprawiać Princeton WordNetu. Jego rozwój zatrzymał się na wersji 3.1., która jest łączona ze Słowosiecią za pomocą relacji międzyjęzykowych. Powstała jednak nowa inicjatywa rozwoju angielskiego wordnetu (English WordNet, https://en-word.net/), w związku z którą użytkownicy Princeton WordNetu mają możliwość zgłaszania uwag na stronie: https://github.com/globalwordnet/english-wordnet).

Łączenie relacjami z Walentego

Łączenie Słowosieci z warstwą semantyczną Walentego odbywa się za pomocą relacji synsetów, oznaczonych jako „WAL", oddających preferencje semantyczne z Walentego, czyli związki roli zachodzące pomiędzy dwoma jednostkami lub ich predefiniowanymi zbiorami (synsetami). Z powodu istnienia w Walentym predefiniowanych zbiorów jednostek (synsetów) i przechodniości preferencji na hiponimy zdecydowano, aby rzutowanie na warstwę semantyczną Walentego odbyło się na poziomie synsetów, nie jednostek.

Pierwsza iteracja łączenia Słowosieci z warstwą semantyczną słownika walencyjnego Walenty miała miejsce w 2019 roku, w związku z tym relacji pomiędzy dwoma zasobami wyrażonymi w formie grafu jest w tej chwili niewiele. Aby mieć do nich dostęp, należy wgrać odpowiedni „konfig" do WordNet Looma. Nie jest on dodawany domyślnie, ponieważ są takie synsety, które mają ogromną liczbę relacji z Walentego (np. PODMIOT). Ich wgrywanie powodowało zawieszanie się aplikacji.

Nie wykluczamy, że w rzutowaniu na warstwę semantyczną Walentego mogą się zdarzyć błędy. Wynikają one z tego, że do opisu semantycznego w Walentym była wykorzystywana Słowosieć w wersji 2.0. (czasownik, który jest najliczniej reprezentowany w Walentym, był rozwijany zwłaszcza w wersji 4.0). Osoby, które pracują w wersji Looma z konfigiem, mogą takie błędy komentować. Komentujemy je ustawiając status: *Błąd* dla całego synsetu i w komentarzu błędu pisząc, że błąd dotyczy relacji WAL.

Anotacja emotywna

Anotacja emotywna (tzw. anotacja sentymentem lub wydźwiękiem) była prowadzona dla wersji 3.0 Słowosieci. Pojawiają się w niej błędy, które najczęściej związane są z tym, że w początkowej fazie anotacji anotatorzy nie mieli odpowiedniego narzędzia do swoich prac (modułu w Loomie), a część z nich nie miała wykształcenia językoznawczego (Słowosieć emo anotowały pary psycholog - lingwista). Proces anotacji Słowosieci uległ „zamrożeniu", niewykluczone, że zostanie ona wznowiona. Póki co w Słowosieci nie pracują anotatorzy sentymentem. Jeśli lingwista Słowosieci zauważy „lżejszy" błąd w anotacji (np. w przykładzie użycia, dodanym przez anotatora), może go sam poprawić. Cięższe błędy, jak nieodpowiednia anotacja, czy problemy wynikające ze złego zrozumienia znaczenia przez anotatora, zostawiamy „na później", dla przyszłego zespołu anotatorów, wpsiując błąd w tabelę: https://docs.google.com/spreadsheets/d/1vODY2j7KQFvamNqdd8stnZRuUxm3ctHZF4Er18MSLpE/edit#gid=0. Oprócz tego znakujemy takie jednostki statusem: *Błąd* i w komentarzu wpisujemy, że błąd dotyczy anotacji emotywnej.