POLONICA ???
PL ISSN 0137-9712

ADAM PRZEPIÓRKOWSKI

Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN

Jedna z siedmiu złotych maksym Geoffrey'a Leecha dotyczących lingwistycznego znakowania korpusów głosi, że systemy znaczników powinny być możliwie niekontrowersyjne i maksymalnie niezależne od jakiejkolwiek teorii lingwistycznej (Leech, 1993). Postulat ten jest w oczywisty sposób nierealistyczny, gdy odnieść go do składniowych schematów znakowania korpusów — każdy tzw. bank drzew, czyli korpus znakowany składniowo, korzysta z jakiejś teorii składni, a przez to jest kontrowersyjny, gdyż nie istnieją niekontrowersyjne teorie składni.

Okazuje się jednak, że także znakowanie morfoskładniowe, tj. znakowanie słów znacznikami zawierającymi informację o klasie gramatycznej danego słowa i o wartościach kategorii gramatycznych takich jak przypadek i rodzaj, wymaga podjęcia szeregu kontrowersyjnych decyzji, w tym decyzji składniowych.

Celem niniejszego artykułu jest omówienie niektórych składniowych uwarunkowań tagsetu stosowanego w projekcie KBN 7 T11C 043 20 (por. Przepiórkowski i in. 2002) i opisanego w artykule Woliński 2002, a także przedstawienie przede wszystkim składniowych zasad dezambiguacji znaczników morfosyntaktycznych (tzw. *tagów*) tego tagsetu.[1] Artykuł ten zakłada znajomość terminologii i rozwiązań wprowadzonych w artykule Woliński 2002.

1. Składniowe uwarunkowania tagsetu IPI PAN

Jednym z celów projektu KBN 7 T11C 043 20 jest stworzenie dużego lingwistycznie anotowanego korpusu języka polskiego, który mógłby być wykorzystywany w różnych zastosowaniach, nie tylko tych najbardziej typowych, leksykograficznych. Ważne jest zatem stworzenie maksymalnie przejrzystego tagsetu i jasnych zasad dezambiguacji, w minimalnym stopniu narzucających użytkownikowi poglądy morfosyntaktyczne czy składniowe twórców korpusu.

Postulat ten był realizowany już w trakcie projektowania tagsetu, gdzie przejrzystość ma zapewnić zestaw klas gramatycznych wyodrębnionych morfoskładniowo, a nie na pod-

[1] Niniejszy artykuł powstał w wyniku wielu dyskusji z udziałem uczestników niniejszego projektu, przede wszystkim Marcina Wolińskiego, Łukasza Dębowskiego i Elżbiety Hajnicz, a także Zygmunta Saloniego, Moniki Korczakowskiej i Beaty Wójtowicz. Zasady dehomonimizacji opisane w niniejszym artykule odpowiadają stanowi projektu KBN 7 T11C 043 20 pod koniec roku 2002 i mogą ulec zmianie w wyniku dalszych prac projektowych.

stawie zbioru kryteriów morfologicznych, składniowych i semantycznych o niejasnych wagach (por. tradycyjne rozumienie pojęć takich, jak *liczebnik* i *zaimek*). Ponadto wydzielone zostały kategorie reprezentujące 'mieszane części mowy' (ang. *mixed categories*), dzięki czemu użytkownik może sam decydować, czy na przykład odsłowniki typu *pisanie* i *picie* traktować jako formy czasownikowe, czy też jako formy rzeczownikowe, czy imiesłowy przymiotnikowe są formami przymiotnikowymi, czy też czasownikowymi itd.

Celem niniejszego punktu jest pokazanie, że stworzenie względnie przejrzystego tagsetu morfosyntaktycznego wymaga także podjęcia licznych decyzji składniowych.

1.1. Fleksemy synkretyczne i defektywne

Zagadnieniem z pogranicza morfologii i składni jest wybór odpowiednich klas gramatycznych (zwanych u nas klasami fleksyjnymi) i decyzja, które zbiory form wyrazowych o charakterze synkretycznym i/lub defektywnym należy wyodrębnić w nowe klasy gramatyczne, a które włączyć do już istniejących klas gramatycznych.

Przykładem fleksemów defektywnych są fleksemy rzeczownikowe *plurale tantum*, np. WUJOSTWO czy NOŻYCE, oraz fleksemy traktowane przez nas jako rzeczowniki *singulare tantum*, np. KTO i co. W obu wypadkach mamy do czynienia ze zbiorami form, których paradygmat różni się od typowego paradygmatu rzeczownikowego, a zatem — traktując serio pojęcie fleksemu i klasy fleksyjnej — powinniśmy wyróżnić klasę różniącą się od klasy rzeczowników, subst, posiadaniem słownikowej, a nie fleksyjnej, kategorii liczby. Do klasy tej należałyby fleksemy *plurale tantum*, o ustalonej liczbie mnogiej, oraz fleksemy typu KTO i co, o ustalonej liczbie pojedynczej. W interesie ograniczenia liczby klas fleksyjnych nie czynimy jednak tego, dopuszczając fleksemy defektywne. Upoważnia nas do tego implicytne kryterium dystrybucyjne, a więc składniowe: fleksemy *plurale tantum* i, w mniejszym stopniu, *singulare tantum* mają dystrybucję podobną do dystrybucji zwykłych rzeczowników, lecz ograniczoną do pozycji dopuszczających rzeczowniki w liczbie mnogiej lub, odpowiednio, w liczbie pojedynczej.

Z innym problemem mamy do czynienia w wypadku fleksemów synkretycznych[2] takich, jak EMU, PEPSI czy WIDZIMISIĘ. Jak pisze Wojdak 2000, „włączenie [leksemów synkretycznych — AP] do klasy wyrazów odmiennych jest możliwe dopiero po uwzględnieniu ich cech składniowych" (s. 236), tj. ich dystrybucji i łączliwości. Zgodnie z pracą Saloni 1974, fleksemy takie traktujemy jako odmienne i, w wypadku fleksemów podanych powyżej, zaliczamy je do klasy gramatycznej subst.

Ciekawym przykładem fleksemów jednocześnie defektywnych i synkretycznych są liczebniki nieokreślone typu DUŻO i TROCHĘ. Mimo skrajnej synkretyczności tych fleksemów i mimo tego, że mogą one występować wyłącznie w pozycjach mianownikowych, biernikowych i tylko niektórych dopełniaczowych (Przepiórkowski, 1999), uznamy je za fleksemy liczebnikowe, znowu ze względu na ich charakterystykę składniową: na przykład w zdaniach typu (1) poniżej, *dużo* zachowuje się tak, jak liczebniki *wiele* i *pięć* w (2a), a nie tak, jak rzeczownik rodzaju nijakiego *stado* w (2b).

(1) Dużo koni było zmęczonych.
(2) a. Wiele / Pięć koni było zmęczonych.
    b. Stado koni było zmęczone / *zmęczonych.

[2] Pojęcie analogiczne do pojęcia *leksemów synkretycznych*, dyskutowanego w pracy Wojdak 2000.

Decyzje podjęte powyżej wydają się zgodne z tradycją saloniowską (Saloni, 1974) i niekontrowersyjne. Formami znacznie trudniejszymi do zaklasyfikowania są formy zaimków osobowych, wysoce nieregularne już poprzez fakt szczątkowej odmiany przez akcentowość (*go*/*jego*) i poprzyimkowość (*jego*/*niego*), lecz także ze względu na to, że nie jest oczywiste, ilu fleksemom te formy odpowiadają — czy istnieje tylko jeden fleksem zaimkowy, odmienny przez osobę, rodzaj i liczbę, bardzo nieregularny i w dużym stopniu synkretyczny, czy też istnieją cztery fleksemy JA, TY, MY, WY o ustalonej liczbie i osobie, odmienne przez przypadek, oraz piąty fleksem, ON, o ustalonej osobie, lecz odmienny przez liczbę, rodzaj i przypadek? W niniejszej wersji tagsetu IPI PAN wybraliśmy — w pełni świadomi arbitralności tego wyboru — rozwiązanie drugie, jako bliższe tradycji i pozwalające nam uniknąć problemu wyboru formy hasłowej fleksemu zawierającego tak różne formy, jak *ja*, *ty*, *my* i *ona*. Kolejny przykład tego typu decyzji, dotyczącej fleksemu SIEBIE, zostanie podany w następnym podpunkcie.

1.2. Tzw. zaimki anaforyczne

Tzw. zaimki anaforyczne *się* i *siebie* przysparzają wiele kłopotu w rygorystycznych opisach polszczyzny, począwszy od decyzji, czy formy *się* i *siebie* uznać za formy jednego, czy też dwóch (f)leksemów.

Czasami *się* określane jest jako nieakcentowana forma zaimka anaforycznego SIEBIE, a zatem moglibyśmy ją opisać za pomocą kategorii akcentowości, podobnie jak opisywane są formy *go*, *mu* itp., uznajemy jednak na podstawie argumentów poniżej, że *się* i *siebie* należą do dwóch różnych fleksemów.

Po pierwsze, podczas gdy *siebie* zawsze pełni funkcję zaimka anaforycznego (zwrotnego lub recyprokalnego), *się* jest formą wielofunkcyjną: oprócz funkcji zwrotnych, jest także elementem konstrukcji bezosobowych typu (3), konstrukcji medialnych typu (4) i czasowników zwrotnych (*reflexiva tantum*) typu (5) (por. Kupść 2000).

(3) Czyta się / *siebie tę książkę z przyjemnością.
(4) Zupa się / *siebie szybko gotuje.
(5) Jan boi się / *siebie Piotra.

Po drugie, mimo iż jako argument niektórych czasowników *się* wydaje się pełnić tę samą funkcję, co *siebie*, por. (6), to w wypadku innych czasowników, tylko forma *siebie* jest dopuszczalna, por. (7).[3]

(6) a. Janek zobaczył się / siebie w lustrze.
    b. Janek umył się / siebie.
(7) a. Janek rozumie siebie / *się.
    b. Janek kocha siebie / %się.

Po trzecie, *siebie*, lecz nie *się*, dopuszcza użycia ‘prawie zwrotne’ (ang. *near-reflexive*), na które w języku angielskim zwrócił uwagę Jackendoff 1992:

(8) W gabinecie figur woskowych, Ringo Starr umył siebie / %się, a figury pozostałych trzech Beatlesów oddał do czyszczenia chemicznego.

[3] W przykładzie (7b) także forma *się* jest poprawna, ale zmienia znaczenie wypowiedzi, co zostało zasygnalizowane znakiem ‘%’.

W przykładzie (8) jedynie forma *siebie* może być rozumiana jako odnosząca się nie bezpośrednio do podmiotu, lecz do podobizny desygnata podmiotu.

Różnice w dystrybucji i funkcji *się* i *siebie* pozwalają nam zaklasyfikować te formy do różnych fleksemów, SIĘ i SIEBIE. Fleksem SIĘ jest kublikiem — fleksemem nieodmiennym, niespójnikowym, nieprzyimkowym, niepredykatywnym, bez kategorii aspektu.

Trudniej jest podać charakterystykę morfosyntaktyczną fleksemu SIEBIE. Jego formy mają tylko trzy wykładniki tekstowe, *siebie*, *sobie* i *sobą*, różniące się przypadkiem: *siebie* jest wykładnikiem formy biernikowej i dopełniaczowej, *sobie* — celownikowej i miejscownikowej, zaś *sobą* — narzędnikowej. Brak zatem form mianownikowej i wołaczowej. Nie jest natomiast jasne, czy fleksem ten odmienny jest przez liczbę i rodzaj. Choć formy fleksemu SIEBIE nie są zróżnicowane ze względu na wartości tych kategorii, to istnieją argumenty składniowe przemawiające za tym, że formy tego fleksemu posiadają wartości tych kategorii.

(9) a. Zobaczył w lustrze [siebie samego].
b. Zobaczyła w lustrze [siebie samą].
c. Zobaczyli w lustrze [siebie samych].
d. Zobaczyły w lustrze [siebie same].

W przykładach (9a–d), ciągi typu *siebie samego*, *siebie samych* itp. wydają się tworzyć frazy (*Kogo zobaczył? Siebie samego.*), których elementem głównym jest *siebie* (*siebie* nie może usunięte bez wpływu na akceptowalność zdania, zaś *samego* itp. — może). Jeżeli tak, to formy *samego*, *samą* itp. prawdopodobnie uzgadniają wartości liczby i rodzaju z podmiotem domyślnym nie bezpośrednio (są na to 'zbyt głęboko'), lecz za pośrednictwem elementu głównego frazy, której są składnikiem, czyli *siebie*.

Ponieważ argument przedstawiony powyżej nie rozwiewa wszystkich wątpliwości, a uznanie, że SIEBIE odmienia się przez liczbę i rodzaj uczyniłoby formy tego fleksemu wysoce niejednoznacznymi (np. segmentowi *siebie* odpowiadałoby $2 \times 2 \times 9 = 36$ form), przyjęliśmy, że fleksem SIEBIE odmienia się jedynie przez przypadek i w ogóle nie posiada kategorii rodzaju i liczby. Jest on jedynym fleksemem o takich własnościach morfoskładniowych, a więc jedynym elementem klasy fleksyjnej siebie.

Wróćmy na koniec do fleksemu SIĘ i zwróćmy uwagę na to, że zjawiska składniowe dostarczają dodatkowego argumentu za znakowaniem segmentów nie dłuższych niż słowa ortograficzne. Wydawałoby się bowiem, że czasowniki zwrotne (*reflexiva tantum*) należałoby znakować w całości, bez wyodrębniania czasownika i formy *się*. Jednak zjawisko znane jako haplologia znacznika zwrotnego *się* (Rappaport, 1998; Kupść, 1999) pokazuje, że takie podejście skazane jest na niepowodzenie, gdyż czasami wymagałoby ono uznania formy *się* za należącą jednocześnie do dwóch lub więcej (czterech w (13)!) znakowanych segmentów. Przykłady poniżej pochodzą z pracy Kupść 1999.

(10) Boję się głośno roześmiać.
(11) Jan stara się golić codziennie rano.
(12) Jest ładna pogoda i Janowi powinno się przyjemnie dziś przechadzać po parku.
(13) Po tych lekach powinno mu się zacząć udawać mniej obawiać spotkać ze znajomymi sprzed wojny.

Przykłady (11)–(12) pokazują ponadto, że nie zawsze możliwe jest jednoznaczne określenie funkcji danego wystąpienia *się*: w (11) jedna forma *się* pełni jednocześnie funkcję *się*

zwrotnego (*golić się*) i *się* jako części czasownika *reflexivum tantum* (*stara się*), zaś w (12) forma *się* jest jednocześnie częścią konstrukcji bezosobowej (*powinno się*) i czasownika *reflexivum tantum* (*przechadzać się*).

1.3. Kategoria zanegowania i tzw. zaimki negatywne

Kategoria zanegowania, właściwa odsłownikom i imiesłowom przymiotnikowym, została wprowadzona ze względu na ciekawe lecz do niedawna słabo opisane zjawisko języka polskiego jakim jest uzgodnienie negacji (ang. *Negative Concord*; por. Przepiórkowski i Kupść 1999 i Przepiórkowski i in. 2002). Zjawisko to polega, w przybliżeniu, na wymaganiu obecności zanegowanego elementu (de-)werbalnego w zdaniu, w którym występuje forma tzw. zaimka negatywnego, np. NIKT, NIC, NIGDY, ŻADEN:

(14) a. Nie przyszedł na żadne party.
b. *Przyszedł na żadne party.
(15) a. Nieprzyjście na żadne party skończyło się dla niego ostracyzmem otoczenia.
b. *Przyjście na żadne party...
(16) a. Ten facet, niesłuchany przez nikogo, ciągle tu przychodzi.
b. *Ten facet, słuchany przez nikogo,...

Zgodnie z tagsetem stosowanym w niniejszym projekcie, formy *nieprzyjście* i *niesłuchany* są zatem formami fleksemów PRZYJŚCIE i SŁUCHANY o wartości zanegowania równej neg. Ogólniej, formy neg charakteryzują się przedrostkiem *nie-* oraz dopuszczaniem zaimków negatywnych.

Warto tu wspomnieć o dwóch swoistych niekonsekwencjach w znakowaniu korpusu IPI PAN kategorią zanegowania. Pierwsza z nich wynika z niespójnych reguł ortograficznych w języku polskim, które każą pisać *nie* łącznie w wypadku zanegowanych odsłowników i imiesłowów przymiotnikowych, ale osobno w wypadku zanegowanych czasowników i imiesłowów przysłówkowych. Ponieważ przyjęliśmy zasadę, według której segmenty (jednostki znakowane) nie mogą być dłuższe od słów ortograficznych, zmuszeni jesteśmy znakować formy czasownikowe takie jak *przyszedł* w (14) niezależnie od poprzedzającej je partykuły negacji, choć pełni ona składniowo dokładnie taką samą rolę, jak przedrostek *nie-* w wypadku imiesłowów przymiotnikowych i, co więcej, istnieją argumenty za tym, że jest faktycznie przedrostkiem czasownikowym (Kupść i Przepiórkowski, 1997). Należałoby zatem znakować ciągi typu *nie przyszedł* jako zanegowane formy czasownika PRZYJŚĆ.

Druga niekonsekwencja w znakowaniu kategorią zanegowania wynika z traktowania tej kategorii jako produktywnej kategorii fleksyjnej. Gdyby wartość neg kategorii zanegowania przyznawać wszystkim formom zawierającym prefiks *nie* i dopuszczającym zaimki negatywne, to należałoby ją także przyznać niektórym formom finitywnym, niektórym przymiotnikom i niektórym przysłówkom (Przepiórkowski i Kupść, 1999):

(17) Ten facet nienawidzi nikomu pomagać.
(18) Ten facet, nigdy z niczego niezadowolony, działa mi na nerwy.
(19) Westchnął niezauważalnie dla nikogo.

2. Dehomonimizacja w korpusie IPI PAN

W poprzednim punkcie pokazaliśmy, że zaprojektowanie morfosyntaktycznego tagsetu dla języka polskiego i zasad znakowania przy użyciu tego tagsetu wymaga rozważenia szeregu zjawisk składniowych, w tym tak nietypowych, jak haplologia *się*. W niniejszym punkcie zajmiemy się zasadami dezambiguacji znaczników morfoskładniowych na podstawie ich kontekstu składniowego.

2.1. Podstawowe pojęcia

Pojęcie *homonimii* różnie jest rozumiane przez różnych autorów[4], stąd też zachodzi konieczność zdefiniowania tego pojęcia na potrzeby niniejszej pracy.

Za podstawową jednostkę unilateralną, której przypisywana jest interpretacja morfosyntaktyczna uznajemy ***segment*** rozumiany jako ***słowo ortograficzne*** ('od spacji do spacji') lub, w uzasadnionych i ściśle zdefiniowanych wypadkach, właściwy podciąg słowa ortograficznego (np. *łgał|eś*, *napisała|by|m*, *chodź|że*, *potrzebował|że|by|ś*, *do|ń*, *polsko|-|niemiecki*). Bilateralną jednostkę składającą się z segmentu i jego interpretacji morfoskładniowej oraz, być może, semantycznej będziemy nazywać ***formą wyrazową***. Przy takim ustaleniu pojęć wstępnych, przez ***homonimię*** będziemy rozumieć identyczność segmentów dwóch lub większej liczby różnych form wyrazowych. Definicja ta jest zgodna z rozumieniem homonimii w pracach Saloni 1996 i Awramiuk 1999 i jest w istocie zawężeniem pojęcia homonimii do pojęcia ***homografii***[5]. A więc za formy homonimiczne uznawać będziemy także na przykład niehomofoniczne formy wyrazowe o wykładnikach tekstowych *cis* i *Dania*.

Należy zauważyć, że taka definicja homonimii nie ograniczna tego zjawiska do sytuacji, w których różna jest charakterystyka semantyczna przypisywana danym segmentom — rozważamy tutaj także homonimie wewnątrzparadygmatyczne, gdzie dane formy wyrazowe różnią się wyłącznie charakterystyką morfoskładniową (np. przypadkiem).

Ponieważ semantyka leży poza zasięgiem naszych rozważań, ograniczymy powyższe pojęcie ***formy wyrazowej*** do pary uporządkowanej składającej się z segmentu i jednoznacznej interpretacji morfosyntaktycznej.[6] ***Dehomonimizację*** będziemy zatem rozumieć jako ***dezambiguację morfosyntaktyczną***, czyli ograniczenie liczby możliwych analiz morfosyntaktycznych poszczególnych segmentów na podstawie kontekstu, w jakim dane segmenty występują.

Dezambiguacja tak rozumiana nie zawsze prowadzi do znalezienia jednoznacznych opisów — często kontekst nie pozwala rozstrzygnąć, która z kilku możliwych interpretacji morfosyntaktycznych jest w danym wypadku właściwa. Za przykład niech posłuży zdanie (20) z pracy Woliński 2002.

(20) Miałem miał.

Zdanie to jest niejednoznaczne już na poziomie segmentacji, gdyż słowo *miałem* może być jednym segmentem, który otrzyma interpretację rzeczownikową, lub może się ono składać z dwóch segmentów, *miał* i *em*. Po odrzuceniu pierwszej segmentacji, odpowiadającej

[4] Przegląd znaczeń terminu *homonimia* w pracach językoznawczych i słownikach zawarty jest w monografiach Awramiuk 1999, rozdz. 1 i Majewska 2002, rozdz. 1.

[5] Chodzi tu o takie rozumienie pojęcia *homografia*, które nie wyklucza jednoczesnej homofonii.

[6] Pojęcie to odpowiada zatem pojęciu *wyraz morfosyntaktyczny* w pracy Bień 2001.

zdaniu poprawnemu najwyżej w sensie eliptycznym (— *Czym miał to posypać? — Miałem miał. Ale posypał piaskiem.*), zdanie jest nadal wieloznaczne, gdyż zarówno segment *Miał*, jak i segment *miał*, może być wykładnikiem formy wyrazowej nominalnej (forma leksemu MIAŁ) lub werbalnej (forma leksemu MIEĆ). Interpretacje te są analogiczne do narzucających się interpretacji zdań następujących:[7]

(21) a. Miałem piasek.
b. Książkęm miał.

Choć wydawać by się mogło, że jedynie interpretacja odpowiadająca (21a) jest właściwa, teoretycznie należy dopuścić także interpretację odpowiadającą (21b) i, co więcej, obie interpretacje są semantycznie równoważne, więc nawet szerszy kontekst nie pozwoli zdecydować, która z tych dwu interpretacji jest 'właściwa'. Bardziej naturalnym przykładem ilustrującym problem wielości interpretacji trudnej do ujednoznacznienia jest (22).

(22) Autobus wyprzedził samochód.

W zdaniu tym oba segmenty interpretowane rzeczownikowo, tj. *Autobus* i *samochód*, mogą być wykładnikami form mianownikowych lub biernikowych i znowu nie sposób na podstawie samego zdania rozstrzygnąć, która interpretacja jest 'właściwa'. Zgodnie z tymi przykładami, dehomonimizację rozumiemy jedynie jako ograniczenie liczby interpretacji do tych, które są dopuszczalne w danym ograniczonym kontekście, a nie jako usunięcie wszystkich interpretacji oprócz tej '*jedynej* właściwej'.[8]

Zauważmy na koniec, że problem dehomonimizacji segmentów nie jest równoważny problemowi dehomonimizacji zdań: zdanie (22) ma tylko dwie spójne interpretacje (albo mianownikowy *Autobus* i biernikowy *samochód*, albo biernikowy *Autobus* i mianownikowy *samochód*), podczas gdy dwie interpretacje segmentu *Autobus* i dwie interpretacje segmentu *samochód* odpowiadają w sumie czterem interpretacjom całego zdania — brakuje bowiem informacji o zależnościach pomiędzy interpretacjami poszczególnych segmentów (mianownikowy *Autobus* wyklucza mianownikowy *samochód* itd.). Choć możliwe jest reprezentowanie takich zależności w dezambiguowanym tekście (Oliva i Petkevič, 2002), ograniczymy się tutaj do problemu dehomonimizacji poszczególnych segmentów, ignorując problem spójnej dehomonimizacji całych zdań.

2.2. Problemy dehomonimizacji wewnątrzfleksemowej

Przez ***dehomonimizację wewnątrzfleksemową*** rozumiemy, poprzez analogię z pojęciami *dehomonimizacja wewnątrzparadygmatyczna* i *dehomonimizacja międzyparadygmatyczna* (Awramiuk, 1999), wybór odpowiedniej interpretacji danego segmentu spośród różnych interpretacji w obrębie danego fleksemu. Problem wyboru odpowiedniej wartości kategorii liczby i przypadka dla danego wystąpienia segmentu *okna* jest typowym problemem dehomonimizacji wewnątrzfleksemowej.

W niniejszym projekcie przyjęto szereg kryteriów, na podstawie których dokonywana jest dehomonimizacja wewnątrzfleksemowa. Przede wszystkim, ze względów praktycznych

[7] Zakładamy, że zdanie (21b) jest poprawne przez analogię do zdań *Długom pisał.* i *Niektóre limeryki... trochęm upolitycznił...* To ostatnie zdanie pochodzi z pracy Świdziński 2001 (zdanie w tekście, a nie przykład lingwistyczny).

[8] Termin *ujednoznacznianie* jest zatem mylący; bardziej właściwy byłby termin *u-nie-tak-wielo-znacznianie*.

przyjęto zasadę rozwiązywania wieloznaczności wyłącznie na podstawie kontekstu nie przekraczającego granic zdania. Wynika z tego, że gdyby w korpusie tekstów wystąpiły zdania typu (22), nie zostałyby one w pełni ujednoznacznione.

Dehomonimizacja dokonywana jest przede wszystkim na podstawie tzw. związków zgody i rządu w obrębie danego zdania. Niekiedy wybór odpowieniej interpretacji spośród form danego fleksemu jest prosty i wymaga tylko podstawowej wiedzy językoznawczej. Na przykład segment *stół* może być interpretowany jako mianownik lub biernik, lecz w zdaniu (23) poprawna jest tylko interpretacja biernikowa, co wynika z faktu, że 1) segment *Hermenegilda* ma wyłącznie interpretację mianownikową, 2) segment *oglądała* interpretowany jest jako forma leksemu OGLĄDAĆ łączącego się z mianownikiem i biernikiem, oraz 3) segmenty *Hermenegilda* i *stół* rozumiane są jako wykładniki form będących składniowymi argumentami formy *oglądała*.

(23) Hermenegilda oglądała stół.

Mniej oczywiste jest, jaka jest interpretacja segmentów *cały* i *dzień* w (24), czy segmentu *smutną* w (25).

(24) Hermenegilda oglądała stół cały dzień.
(25) Pamiętam ją smutną.

W tym pierwszym wypadku należy zauważyć, że okoliczniki czasu tego typu występują w bierniku, por. (24′), oraz że przymiotniki predykatywne mogą uzgadniać przypadek z frazą, do której się odnoszą, lub występować w narzędniku, por. (25′).

(24′) Hermenegilda oglądała stół całą godzinę.
(25′) a. Pamiętam go smutnego.
b. Pamiętam go smutnym.

A zatem zdanie (25) jest faktycznie dwuznaczne — segment *smutną* może być interpretowany jako wykładnik formy biernikowej lub narzędnikowej.

W niektórych wypadkach, dehomonimizacja wymaga nie tylko głębokiej świadomości zjawisk składniowych, ale także podjęcia — lub też świadomego uniknięcia — pewnych decyzji teoretycznych. Kilka takich wypadków opisujemy poniżej.

2.2.1. Przypadek liczebników w pozycji podmiotu

Składnia fraz liczebnikowych, zarówno ich wewnętrzna struktura, jak i ich cechy morfoskładniowe widoczne na zewnątrz, jest jednym z najszerzej dyskutowanych zagadnień polskiej składni.[9] Jedna z kontrowersji dotyczy przypadka fraz liczebnikowych w pozycji podmiotu — co najmniej od czasu lwowskiej gramatyki Małeckiego z roku 1863 pojawiają się opinie, że przypadek liczebnika w zdaniach takich, jak (26) to biernik, a nie mianownik.

(26) Pięciu facetów przyszło.

Pełna argumentacja za korzyściami płynącymi z takiego rozwiązania przedstawiona jest w pracy Przepiórkowski 1999 (i powtórzona w pracy Przepiórkowski i in. 2002) i można ją streścić następująco: przy założeniu, że przypadek liczebników w pozycji podmiotu nie

[9] Por. np. Szober 1920, 1922, 1928, Saloni 1976a, 1977, Gruszczyński i Saloni 1978, Wiśniewski 1990, Kopcińska 1992, 1997, Franks 1994, Mieczkowska 1995, Przepiórkowski 1996, 1999, Chachulska 2000, Przepiórkowski i in. 2002.

zależy od rodzaju, tj. jest taki sam dla wszystkich rodzajów gramatycznych, tylko biernikowy liczebnik pozwala prosto wyjaśnić fakty w (27).

(27) a. Tych / Te pięć kobiet przyszło.
b. Tych / *Ci pięciu mężczyzn przyszło.

Jak widać na przykładzie (27a), formy przymiotnika (tradycyjnie: zaimka) ci uzgadniają się albo z liczebnikiem (*te pięć*), albo z rzeczownikiem (*tych kobiet*). Gdyby liczebnik był w mianowniku, nieoczekiwana by była niegramatyczność zdania (27b) z mianownikową formą *ci*. Nieoczekiwany byłby także brak uzgodnienia podmiotu z czasownikiem. Jeżeli jednak przyjąć, że liczebnik występuje w takich konstrukcjach w bierniku, to fakt, że *pięciu mężczyzn* może być modyfikowane jedynie przez formę *tych* wynika z synkretyzmu tej formy pomiędzy biernikiem (uzgodnienie z liczebnikiem) i dopełniaczem (uzgodnienie z rzeczownikiem). Dodatkowo nie wymaga wyjeśniania brak uzgodnienia z czasownikiem poza skonstatowaniem faktu, że język polski jest podobny do innych języków indoeuropejskich, w których czasownik uzgadnia rodzaj, osobę i liczbę jedynie z podmiotami mianownikowymi.

Jedynym znanym nam empirycznym argumentem przeciwko bernikowym podmiotom liczebnikowym jest obserwacja, że w konstrukcjach współrzędnie złożonych w pozycji podmiotu fraza liczebnikowa może występować obok rzeczownikowej frazy mianownikowej, a zatem fraza liczebnikowa również musi być mianownikowa (Saloni, 1976a), np.:

(28) Do kina poszło [[pięciu facetów] i [ich bracia]-NOM].

Argument ten jest jednak tak mocny, jak założenie, że tylko frazy o tej samej wartości przypadka mogą być współrzędnie złożone. Łatwo pokazać, że założenie to jest nieprawdziwe:

(29) a. Dajcie [wina-GEN i [całą świnię]-ACC]!
b. Zrobię to [[późnym wieczorem]-INS lub [następnego ranka]-GEN].
c. [Kto-NOM, co-ACC i komu-DAT] dał?

Wydaje się zatem, że istnieje mocny argument za biernikiem liczebnika w pozycji podmiotu i nie ma empirycznych argumentów przeciw takiemu rozwiązaniu, należałoby więc takie formy liczebnikowe znakować jako biernikowe. Jednak jak powiedziano powyżej, w niniejszym projekcie uznajemy kryterium użyteczności i przejrzystości za nadrzędne, a ono nakazuje nam znakować takie formy jako mianownikowe. Rozumowanie które prowadzi do takiego wniosku jest następujące: wobec wciąż żywej kontrowersji dotyczącej przypadka liczebnika w pozycji podmiotu, formy liczebnikowe należy znakować tak, żeby nie wykluczać żadnego z tych stanowisk. Gdyby były one znakowane jako biernikowe, trudne by było odtworzenie przez użytkownika korpusu informacji, które z biernikowych form liczebnikowych znajdują się w 'pozycji mianownikowej' — reprezentowany byłby zatem tylko jeden pogląd. Jeżeli jednak takie formy będą znakowane jako mianownikowe, reprezentowane będą oba poglądy, gdyż użytkownicy przekonani, że są to w istocie formy biernikowe będą mogli bez trudu zastąpić wszystkie znaczniki typu num:pl:nom:... znacznikami typu num:pl:acc:....

Zgodnie z powyższym, w dalszej części niniejszego artykułu przyjmujemy, że formy liczebnikowe w pozycji podmiotu są w istocie formami mianownikowymi.

2.2.2. Rząd i uzgodnienie wewnątrz fraz liczebnikowych

Innym ciekawym zjawiskiem dotyczącym składni fraz liczebnikowych jest różnica pomiędzy, z jednej strony, frazami liczebnikowymi w pozycji mianownika i biernika, gdzie liczebnik zwykle łączy się z rzeczownikiem w dopełniaczu, np. (30), a z drugiej strony, frazami liczebnikowymi w celowniku, narzędniku i miejscowniku, gdzie zarówno liczebnik, jak i rzeczownik występują w tym przypadku, por. (31).

(30) a. Pięć-NOM kobiet-GEN przyszło.
b. Widzę pięć-ACC kobiet-GEN.
(31) a. Dałem to pięciu-DAT kobietom-DAT.
b. Rozmawiałem z pięcioma-INS kobietami-INS.
c. Rozmawiałem o pięciu-LOC kobietach-LOC.

Wydawać by się mogło, że jest to zjawisko jedynie składniowe, zależące od pozycji składniowej danej frazy i nie wymagające uwzględnienia na poziomie systemu znaczników morfosyntaktycznych, lecz podobną chwiejność widać także w mianownikowych frazach liczebnikowych, których elementami głównymi są męskoosobowe formy 'niskich' liczebników (ang. *paucal numerals*) DWA, TRZY i CZTERY:

(32) a. Dwaj-NOM faceci-NOM przyszli.
b. Dwóch-NOM facetów-GEN przyszło.

Przy założeniu, które uczyniliśmy w poprzednim punkcie, że forma *dwóch* w pozycji podmiotu jest formą mianownikową, tradycyjne kategorie rodzaju, liczby i przypadka nie pozwalają odróżnić form *dwaj* i *dwóch* — zajmują one tę samą klatkę paradygmatu liczebnika DWA. Z tego powodu Bień i Saloni (1982) wprowadzili kategorię akomodacyjności o wartości rec dla form takich, jak mianownikowe *dwóch*, występujące z rzeczownikiem niemianownikowym, i congr dla form takich, jak mianownikowe *dwaj*, występujące z rzeczownikiem w mianowniku.

W obecnym tagsecie kategoria ta została rozciągnięta na wszystkie formy wszystkich liczebników, dzięki czemu mianownikowe i biernikowe formy liczebnikowe w (30) znakowane są jako rec (w odpowiedniej pozycji znacznika), zaś celownikowe, narzędnikowe i miejscownikowe formy liczebnikowe w (31) — jako congr. Pozostaje jednak pytanie, jak znakować dopełniaczowe formy liczebnikowe.

(33) a. Nie widziałem pięciu kobiet.
b. Bałem się pięciu kobiet.

Narzucająca się interpretacja wartości kategorii akomodacyjności, chyba zgodna z intencjami autorów artykułu Bień i Saloni 1982, jest taka, że formy rec to formy rządzące dopełniaczem, zaś formy congr to formy uzgadniające przypadek. Czy w wypadku liczebników w dopełniaczu mamy jednak do czynienia z rządem dopełniacza, czy też z uzgodnieniem dopełniacza?

Najprościej byłoby przyjąć któreś z dwóch możliwych rozwiązań arbitralnych i założyć, że wszystkie dopełniaczowe formy liczebnikowe mają wartość akomodacyjności równą rec, lub też — że wszystkie mają wartość congr. W pracy Przepiórkowski 1999 podano, na podstawie dystrybucji defektywnych liczebników nieokreślonych typu *dużo* i *trochę*, argumenty (powtórzone w Przepiórkowski i in. 2002) za trzecim stanowiskiem, mianowicie takim, że

podział na rec i congr przebiega w poprzek wystąpień dopełniaczowych form liczebnika: formy liczebnikowe w pozycji przyczasownikowego (i przyprzyimkowego) dopełniacza strukturalnego rządzą dopełniaczem, zaś w innych pozycjach — uzgadniają przypadek z rzeczownikiem.

Jest to kolejny przykład kontrowersji składniowej, w której twórcy tagsetu nie powinni opowiadać się po żadnej ze stron. Z tego powodu w obecnej wersji tagsetu i zasad dehomonimizacji przyjęto, że dopełniaczowe formy liczebnikowe pozostają niejednoznaczne co do wartości akomodacyjności, dzięki czemu będzie je można ujednoznacznić na dalszych etapach przetwarzania korpusu w zależności od wybranej teorii składniowej.

2.2.3. Dehomonimizacja rodzaju

Już sama liczba rodzajów w języku polskim jest kontrowersyjna. Oprócz naiwnego podejścia, według którego rodzaje są trzy (męski, żeński, nijaki), funkcjonuje podejście „szkolne" (wspomniane powyżej 3 rodzaje w liczbie pojedynczej i dwa w mnogiej — męskoosobowy i niemęskoosobowy), 5 rodzajów Mańczaka 1956 i 9 rodzajów Saloniego 1976b, będących uszczegółowieniem propozycji Mańczaka i wyodrębnionych na podstawie następujących kontekstów:

(34) m1. Widzę jednego albo dwóch spośród tych ____, których lubię.
m2. Widzę jednego albo dwa spośród tych ____, które lubię.
m3. Widzę jeden albo dwa spośród tych ____, które lubię.
n1. Widzę jedno albo dwoje spośród tych ____, które lubię.
n2. Widzę jedno albo dwa spośród tych ____, które lubię.
f. Widzę jedną albo dwie spośród tych ____, które lubię.
p1. Widzę jedno albo dwoje spośród tych ____, których lubię.
p2. Widzę jedne albo dwoje spośród tych ____, które lubię.
p3. Widzę (jedną albo dwie pary) spośród tych ____, które lubię.

Ponadto, Przepiórkowski i in. 2002 i Woliński 2001 niezależnie proponują ograniczenie liczby rodzajów w języku polskim przy jednoczesnym zdaniu sprawy z różnic pomiędzy kontekstami n1 i n2, oraz pomiędzy rodzajami *plurale tantum* p1, p2 i p3.

Obecnie przyjmujemy w niniejszym projekcie saloniowską klasyfikację rodzajów w języku polskim przedstawioną w (34), jako najbardziej szczegółową, i zakładamy, że kategoria rodzaju ma ten sam zestaw wartości dla wszystkich klas gramatycznych, dla których jest właściwa. Oznacza to jednak istnienie dużej liczby synkretyzmów — na przykład segment *zrobiły* ma siedem różnych interpretacji rodzajowych. Potrzebne są zatem jasne reguły dehomonimizacji takich wieloznaczności.

Podstawowym kryterium dehomonimizacji rodzaju jest uzgodnienie: zakładamy, że przymiotniki, imiesłowy przymiotnikowe oraz pseudoimiesłowy uzgadniają rodzaj z rzeczownikiem, z którym się łączą. Wyjątkiem od tej reguły jest uzgodnienie z frazą nominalną złożoną współrzędnie, co spowodowane jest faktem, że frazy takie mają dosyć skomplikowaną składnię zewnętrzną, nie dającą się ująć w kilka prostych reguł:[10]

(35) a. ... patriotyzm i poczucie obowiązku wobec kraju nakazywały... (A1024)
b. ... lekkość i odporność polimerów na wpływy atmosferyczne predestynują je do powyższych zastosowań. (C0138)

[10] Przykłady poniżej pochodzą z Korpusu *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz i in., 1990) i oznaczone są numerem odpowiedniej próbki.

(36) a. Ogólnym dążeniem... jest pogłębienie i rozszerzenie pozytywnych tendencji... (A0810)
b. ... zmniejszyła się produkcja i skup żywca... (A0999)

(37) a. Opanowanie i rozwinięcie technologii sklepienia dało więc początek... (C0134)
b. Spienianie i utwardzanie wypełniacza... wymaga nagrzania formy... (C0206)

Jak pokazują przykłady powyżej, już uzgodnienie liczby wymyka się prostym regułom składniowym: w przykładach (35) mamy do czynienia z sytuacją, gdy fraza współrzędnie złożona ze spójnikiem *i* i podrzędnymi frazami nominalnymi w liczbie pojedynczej uzgadnia się z czasownikiem w liczbie mnogiej. W przykładach (36) podobne frazy współrzędnie złożone uzgadniają się natomiast z formami czasownikowymi w liczbie pojedynczej — prawdopodobnie są to przykłady uzgodnienia do najbliższego składnika konstrukcji współrzędnie złożonej, opisywanego m.in. w pracach Kallas 1974, 1993.

Takie uzgodnienie do najbliższego składnika jest jednak typowe dla fraz współrzędnie złożonych występujących na prawo od uzgadnianej formy czasownikowej, a nie na lewo:

(38) a. Na brzegu leżała łódź i wiosło.
b.?*Łódź i wiosło leżało na brzegu.

Z tego wynika, że w wypadku przykładów (37), w których fraza złożona także uzgadnia się z formą czasownika w liczbie pojedynczej, mamy prawdopodobnie do czynienia z innym zjawiskiem, polegającym na 'przeniesieniu znaczenia' (ang. *reference transfer*; Nunberg 1978). Zjawisko to, na pograniczu składni, semantyki i pragmatyki, omawiane jest w odniesieniu do podmiotowych konstrukcji współrzędnie złożonych m.in. w pracach Morgan 1972 i Pollard i Sag 1994, na podstawie angielskich przykładów takich, jak (39).

(39) Doing phonology problems and drinking vodka makes me sick.

Semantyczność reguł uzgodnienia z frazą współrzędnie złożoną w języku polskim dobitnie ilustrują także przykłady (40) poniżej, dotyczące właśnie kategorii rodzaju.[11]

(40) a. Kobieta i dziecko przyszli / ?*przyszły.
b. Kobieta i dziewczę przyszły / ?*przyszli.

Składniki frazy współrzędnie złożonej w (40a) mają identyczne wartości wszystkich kategorii morfoskładniowych, co odpowiadające im składniki w (40b), w szczególności *dziecko* i *dziewczę* mają tę samą wartość kategorii rodzaju, `n1`, a jednak — z przyczyn semantycznych — inaczej uzgadniają się z formą finitywną czasownika. Z powodu tych i innych problemów związanych ze składnią zewnętrzną fraz współrzędnie złożonych, w niniejszym projekcie przyjęto zasadę ignorowania w procesie dehomonimizacji wewnątrzfleksemowej ewentualnych związków zgody z takimi frazami.

2.3. Problemy dehomonimizacji międzyfleksemowej

W poprzednim punkcie omawialiśmy niektóre problemy związane z dehomonimizacją wewnątrzfleksemową. Tutaj zilustrujemy problem wyboru odpowiedniego fleksemu spośród tych fleksemów, do których mogą należeć interpretacje danego segmentu.

Dehomonimizacja międzyfleksemowa w większym stopniu niż dehomonimizacja wewnątrzfleksemowa odwołuje się do intuicji semantycznych. Często stosowanym, choć chyba

[11] Na przykłady takie zwracają uwagę Przepiórkowski i in. 2002.

rzadko formułowanym eksplicite kryterium wyboru odpowiedniego fleksemu dla danego segmentu jest test, który nazwiemy ***kryterium fleksyjnej niezmienności znaczenia***. Zilustrujmy to kryterium na prostym przykładzie (41).

(41) Nie lubię gotować, ale lubię piec, szczególnie ciasta.

Jest oczywistym, że w przykładzie tym interpretacja segmentu *piec* jest czasownikowa, a ściślej, inf:imperf, a nie subst:sg:nom.acc:m3. Dlaczego jest to oczywiste? Nie decyduje o tym składnia, gdyż składniowo zdanie powyższe też jest niejednoznaczne i teoretycznie mogłoby mieć strukturę analogiczną do struktury zdania (42).

(42) Nie lubię gotować, ale lubię różne potrawy, szczególnie ciasta.

Decyduje o tym znaczenie zdania (41), a szczególnie znaczenie słowa *piec* pozostające w systematycznej relacji do odpowiednich znaczeń słów *piekę* i *pieczenie*, a nie słów *piece* i *piecami*.

Spróbujmy sprecyzować to kryterium. Załóżmy, że wśród teoretycznie możliwych (zadanych słownikowo) interpretacji segmentu s jest interpretacja mówiąca, że s to wykładnik tekstowy formy f należącej do fleksemu F. Kryterium fleksyjnej niezmienności znaczenia mówi, że segment s może być interpretowany w pewnym kontekście K jako forma f fleksemu F tylko wtedy, gdy dla każdej formy f' tego fleksemu istnieje kontekst K', w którym f' ma to samo znaczenie, co f w kontekście K, uwzględniając regularne różnice znaczeń wynikające z innych wartości kategorii gramatycznych tych form (np. różnice znaczeń wynikające z innej wartości kategorii liczby).

Według powyższego kryterium, segment *piec* ze zdania (41) nie może być interpretowany jako subst:sg:nom.acc:m3, tj. jako należący do fleksemu rzeczownikowego, gdyż istnieją takie formy tego fleksemu (np. *piecem*), dla których nie da się znaleźć kontekstów, w których znaczyłyby one to samo, co *piec* w (41). Kryterium to jest stosowane m.in. we wstępie do *Słownika frekwencyjnego polszczyzny współczesnej* (SFPW; Kurcz i in. 1990):

> Za słowoformy [tj., w naszym rozumieniu, segmenty — AP] rzeczownikowe uznano te jednostki, które można uważać za nie wchodzące do opozycji rodzajowej obejmującej w mianowniku liczby pojedynczej rodzaj nijaki obok męskiego i żeńskiego — przy zachowaniu tego samego znaczenia. Stosując to kryterium, uznano nie tylko istnienie haseł rzeczownikowych takich, jak np.: LEŚNICZY (rzeczowniki rodzaju męskiego), WYTYCZNA (rzeczownik rodzaju żeńskiego), lecz również...: POŚPIESZNY (rzeczownik rodzaju męskiego), PRZEWODNICZĄCY (rzeczownik rodzaju męskiego), PRZEWODNICZĄCA (rzeczownik rodzaju żeńskiego)... Uznawano więc istnienie leksemów rzeczownikowych o ustalonym rodzaju obok leksemów przymiotnikowych odmiennych przez rodzaj, np. leksemu ŚREDNIA rodzaju żeńskiego obok leksemu przymiotnikowego ŚREDNI.

A zatem, dla segmentów takich, jak *pośpieszny* (s) w zdaniu „Pośpieszny wjechał na stację" (K), wykluczona zostanie interpretacja tego segmentu jako mianownikowej formy przymiotnikowej (f) należącej do przymiotnikowego fleksemu POŚPIESZNY (F), gdyż forma nijaka tego fleksemu (f') nie może w żadnym kontekście (K') mieć tego samego znaczenia, co *pośpieszny* w tym zdaniu.

Oczywiście kryterium to, choć podane tutaj w sposób eksplicytny, nie jest w pełni ścisłe, gdyż posługuje się tak nieścisłymi pojęciami, jak *znaczenie*, szczególnie *to samo*

znaczenie, uwzględniają regularne różnice znaczeń wynikające z innych wartości kategorii gramatycznych. Wydaje się jednak, że mimo tych nieścisłości, kryterium to jest pożyteczne, gdyż pozwala na wyeliminowanie pewnych interpretacji.

Poniżej omawiamy kilka szczegółowych problemów dotyczących dehomonimizacji międzyfleksemowej, w wypadku których kryterium fleksyjnej niezmienności znaczenia okazało się niewystarczające. Problemy te mają dwa główne źródła: brak wyraźnych różnic znaczeniowych pomiędzy fleksemami zawierającymi homofoniczne formy i nieodmienność takich fleksemów, uniemożliwiająca zastosowanie powyższego kryterium.

Rzeczownik czy przymiotnik: *wszystko*, *to*, *tamto* Formy należące do fleksemów rzeczownikowych rodzaju n2: WSZYSTKO, TO i TAMTO są homonimiczne z formami fleksemów przymiotnikowych, odpowiednio, WSZYSTEK, TEN i TAMTEN. Fleksemy te mają na tyle zbliżone znaczenia, że nie jest możliwe stosowanie kryterium fleksyjnej niezmienności znaczenia.

W wypadku, gdy można określić rodzaj danej formy tego typu i nie jest to n2, forma ta należy do odpowiedniego fleksemu przymiotnikowego (np. *Znam je w s z y s t k i e.*, *Zaproś t y c h, których znasz.*). W pozostałych wypadkach decyzja zależy od pozycji danej formy w zdaniu: jeżeli występuje ona w pozycji typowo rzeczownikowej, np. jako argument czasownika lub przyimka, uznajemy ją za formę rzeczownikową, np. *Dokonała t e g o sama.*, *przede w s z y s t k i m* itp. W przeciwnym wypadku uznajemy daną formę za formę przymiotnikową, np. *T e g o wyczynu dokonała sama.* itp.

W wypadku połączeń typu *to wszystko* pełniących funkcję fraz rzeczownikowych (np. *Już t o w s z y s t k o widziałem.*), forma *to* jest arbitralnie traktowana jako forma przymiotnikowa, zaś *wszystko* — rzeczownikowa. Rozwiązanie to jest wzorowane na SFPW.

Rzeczownik czy kublik: *czym* Podobne rozwiązanie stosujemy w wypadku segmentu *czym*, który może zostać uznany za rzeczownik (subst) lub kublik (qub):

(43) subst: *przy czym*;
(44) qub: *czym prędzej*.

Gdy *czym* występuje w pozycji typowo rzeczownikowej, np. jako argument czasownika łączącego się z frazą nominalną, lub jako argument przyimka, uznajemy, że jest to segment rzeczownikowy. W przeciwnym wypadku, nadajemu mu interpretację kublikową.

Rzeczownik, przymiotnik, predykatyw, spójnik czy kublik: *to* Segment *to* może zostać uznany za rzeczownik (subst), przymiotnik (należący do fleksemu TEN; adj), spójnik (conj), predykatyw (pred) lub kublik (qub).

Forma odmienna *to* została omówiona powyżej: jest formą rzeczownikową, gdy zajmuje pozycję przyczasownikową (por. *Jan zabrał t o*) lub przyprzyimkową (*Jan popatrzył na t o*), oraz formą przymiotnikową, gdy zajmuje pozycję przyrzeczownikową (por. *Jan zabrał t o pudełko*). Zgodnie z tą regułą, we frazie *przy t y m*, segment *tym* jest formą rzeczownikową (np. *Upierał się przy t y m.*), o ile nie występuje po nim forma rzeczownikowa modyfikowana przez *tym* (np. *Upierał się przy t y m pomyśle.*).

Wydawać by się mogło, że rzeczownikowe i przymiotnikowe formy *to*, tradycyjnie nazywane zaimkami, powinno być łatwo odróżnić od *to* predykatywnego, spójnikowego czy partykuło-przysłówkowego (tzw. kublika). Okazuje się jednak, że odróżnienie predykatywnej i rzeczownikowej formy *to* może sprawiać kłopoty. Przyjmujemy, że forma

to jest predykatywem w przykładach typu *To (są/były) bliźniaki.* czy *Janek to drań.*, zaś rzeczownikiem w konstrukcjach typu *To było stołem.* (np. wskazując na kilka kawałków drewna). Uznajemy zatem, że *to* jest predykatywem w konstrukcjach typu *to (jest/było/będzie) NP*, gdy fraza rzeczownikowa NP jest mianowniku, zaś rzeczownikiem, gdy fraza ta jest w narzędniku.

Nieodmienne *to* będące składnikiem spójników złożonych *choć..., to...*; *jak..., to...*; *jeżeli..., to...*; *skoro..., to...* uznajemy za spójnik. *To* bywa także spójnikiem samodzielnym (np. *Chcesz, to kładź się spać.*; *Nawarzyłeś piwa, to teraz je wypij.*).

Ponadto nieodmienne *to* jest kublikiem, jeśli pełni rolę wzamacniającą i po jego usunięciu sens wypowiedzenia nie ulega zmianie, np. *w którym to zakładzie*, *coraz to chłodniej* itp.

Podobne kryteria można sformułować dla segmentu *co*, który w zależności od kontekstu może być uznany za rzeczownik (subst, rodzaj n2), spójnik (conj), kublik (qub) lub przyimek (prep:acc).

Forma finitywna czy kublik: *może* Forma *może* jest kublikiem w zwrotach takich, jak *On może przyjdzie.*, natomiast formą nieprzeszłą w przykładach *On może przyjść.*. W niektórych sytuacjach trudno jest te dwie formy odróżnić, gdyż obie mogą mieć znaczenie epistemicznego operatora modalnego. Gdy usunięcie formy *może* powoduje nieakceptowalność zdania, to przyjmujemy, że jest to forma nieprzeszła, np. (45).

(45) a. Maria *(może) to zignorować.
b. Maria być *(może) przyjdzie.

Gdy segment *może* jest jednoznacznie rozumiany jako modyfikujący zdanie, którego elementem głównym jest jest forma finitywna, to jest on formą kublika MOŻE:

(46) Maria może przyjdzie.

W niektórych wypadkach testy te jednak nie wystarczają do podjęcia decyzji, jak choćby w przykładzie (47), w którym dowolny z segmentów *może* może być interpretowany jako forma nieprzeszła, poprzez analogię z (48).

(47) Może może przyjść.
(48) a. Może mógłby przyjść.
b. Mógłby może przyjść.

Spójnik czy kublik: *czy*, *skąd*, *gdzie*, *kiedy* itp. Nieodmienne wyrazy pytajne *czemu*, *czy*, *dlaczego*, *dokąd*, *gdzie*, *kiedy*, *skąd* itp. występujące na poziomie zdania głównego uznajemy za kubliki, np. (49). Te same formy w funkcji wyrazów względnych lub wyrazów pytajnych wprowadzających zdanie podrzędne uznajemy za spójniki, np. (50).

(49) a. Czy Janek przyszedł?
b. Skąd to masz?
c. Gdzie mówiłeś, że masz się z nim spotkać?
d. A Maria przyjdzie kiedy?

(50) a. Zastanawiałam się, czy to dobry przykład.
b. Zapytałem go, skąd to wie.
c. To jest to miejsce, gdzie go spotkałem.
d. Mówiłeś, gdzie go spotkasz?
e. Zapłacę ci, kiedy to zrobisz.

Czasami kryteria te nie wystarczają do nadania odpowiedniej interpretacji takim formom. Na przykład nie jest jasne, czy w przykładzie (51) segment *czy* jest formą spójnikową, jak w (52a), czy też formą kublika, jak w (52b).

(51) Czy przyjdzie — nie wiem.
(52) a. Nie wiem, czy przyjdzie.
b. Czy przyjdzie? Nie wiem.

Forma *czy* jest także spójnikiem łączącym zdania równorzędne, w tej funkcji często równoważny *lub* i *albo*, np.: *drzemał c z y udawał, że drzemie*, *był smutny c z y też znudzony*, *tak c z y inaczej*.

3. Podsumowanie

Celem niniejszego artykułu było pokazanie, że znakowanie morfosyntaktyczne wymaga podjęcia szeregu nietrywialnych decyzji składniowych, zarówno na etapie projektowania tagsetu, jak i w trakcie dehomonimizacji wewnątrzfleksemowej i międzyfleksemowej.

Wydawałoby się, że takie decyzje powinny powodować mniejszą przejrzystość systemu znaczników morfosyntaktycznych i większą zależność tego systemu od danej teorii składniowej, wbrew maksymom Geoffrey'a Leecha. W artykule tym próbowaliśmy jednak pokazać, że jest wręcz przeciwnie — to właśnie dzięki znajomości alternatywnych opisów składniowych danego zjawiska i świadomości arbitralności wielu rozwiązań można zaprojektować system znaczników morfoskładniowych i reguły ich ujednoznaczniania w taki sposób, by żadego z tych alternatywnych poglądów na składnię nie wykluczać.

Mamy nadzieję, że w projekcie opisanym w niniejszej serii artykułów udało nam się zbliżyć do ideału szczegółowego, precyzyjnego i jednocześnie precyzyjnego tagsetu dla języka polskiego.

Bibliografia

Awramiuk, Elżbieta. (1999) *Systemowość polskiej homonimii międzyparadygmatycznej*. Białystok: Wydawnictwo Uniwersytetu w Białymstoku.

Bień, Janusz S. (2001) „O pojęciu wyrazu morfologicznego". Gruszczyński i in. (2001), 67–78.

Bień, Janusz S. i Zygmunt Saloni. (1982) „Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)". *Prace Filologiczne* XXXI: 31–45.

Borsley, Robert D. i Adam Przepiórkowski, red. (1999) *Slavic in Head-driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications.

Chachulska, Beata. (2000) „Jeszcze o liczebnikach typu *pięć*". *Polonica* XX: 137–159.

Franks, Steven. (1994) „Parametric properties of numeral phrases in Slavic". *Natural Language and Linguistic Theory* 12(4): 597–674.

Gruszczyński, Włodzimierz, Urszula Andrejewicz, Mirosław Bańko i Dorota Kopcińska, red. (2001) *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15000 dni pracy naukowej*. Białystok: Wydawnictwo Uniwersytetu Białostockiego.

Gruszczyński, Włodzimierz i Zygmunt Saloni. (1978) „Składnia grup liczebnikowych we współczesnym języku polskim". *Studia Gramatyczne* II: 17–42.

Jackendoff, Ray. (1992) „Madame Tussaud meets the Binding Theory". *Natural Language and Linguistic Theory* 10: 1–31.

Kallas, Krystyna. (1974) „O zdaniach *Pachniał wiatr i morze.*, *Andrzej i Amelia milczeli.*". *Studia z filologii polskiej i słowiańskiej* XIV: 57–71.

———. (1993) *Składnia współczesnych polskich konstrukcji współrzędnych*. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Kopcińska, Dorota. (1992) „Prosta grupa liczebnikowa jako składnik podmiotowej frazy mianownikowej". *Studia Gramatyczne* X: 19–35.

———. (1997) *Strukturalny opis składniowy zdań z podmiotem-mianownikiem we współczesnej polszczyźnie*. Warszawa: Dom Wydawniczy Elipsa.

Kupść, Anna. (1999) „Haplology of the Polish reflexive marker". Borsley i Przepiórkowski (1999), 91–124.

———. (2000) „Lexical analysis of Polish multifunctional reflexive marker". Tracy Holloway King i Irina Sekierina, red., *Annual workshop on formal approaches to Slavic linguistics: The Philadelphia meeting 1999*. Ann Arbor, 214–237.

Kupść, Anna i Adam Przepiórkowski. (1997) „Morphological aspects of verbal negation in Polish". *Proceedings of the second European conference on Formal Description of Slavic Languages, Potsdam, Germany, November 20–22, 1997*. W druku.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran i Jerzy Woronczak. (1990) *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Leech, Geoffrey. (1993) „Corpus annotation schemes". *Literary and Linguistic Computing* 8(4): 275–281.

Mańczak, Witold. (1956) „Ile jest rodzajów w polskim?". *Język Polski* XXXVI(2): 116–121.

Majewska, Małgorzata. (2002) *Homonimia i homonimy w opisie językoznawczym*. Warszawa: Dom Wydawniczy Elipsa.

Mieczkowska, Halina. (1995) *Studia nad liczebnikiem*. Kraków: Universitas.

Morgan, Jerry. (1972) „Verb agreement as a rule of English". Paul Peranteau *et al.*, red., *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Nunberg, Geoffrey. (1978) *The pragmatics of reference*. Ph. D. dissertation, City University of New York.

Oliva, Karel i Vladimir Petkevič. (2002) „Morphological and syntactic tagging of Slavonic languages". Wykład w ramach doktoranckiej szkoły letniej *Empirical Linguistics and Natural Language Processing*, Sozopol, wrzesień 2002.

Pollard, Carl i Ivan A. Sag. (1994) *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press / CSLI Publications.

Przepiórkowski, Adam. (1996) „Case assignment in Polish: Towards an HPSG analysis".

Claire Grover i Enric Vallduví, red., *Studies in HPSG*, tom 12 serii *Edinburgh Working Papers in Cognitive Science*. Centre for Cognitive Science, University of Edinburgh, 191–228.

———. (1999) *Case assignment and the complement-adjunct dichotomy: A non-configurational constraint-based approach*. Rozprawa doktorska, Universität Tübingen.

Przepiórkowski, Adam, Piotr Bański, Łukasz Dębowski, Elżbieta Hajnicz i Marcin Woliński. (2002) „Konstrukcja korpusu IPI PAN". Złożone do *Poloników*.

Przepiórkowski, Adam i Anna Kupść. (1999) „Eventuality negation and negative concord in Polish and Italian". Borsley i Przepiórkowski (1999), 211–246.

Przepiórkowski, Adam, Anna Kupść, Małgorzata Marciniak i Agnieszka Mykowiecka. (2002) *Formalny opis języka polskiego: Teoria i implementacja*. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Rappaport, Gilbert C. (1998) „Clitics as features: A non-semiotic approach". Robert A. Maguire i Alan Timberlake, red., *American contributions to the twelfth international congress of Slavists*. Bloomington, IN: Slavica Publishers, 460–478.

Saloni, Zygmunt. (1974) „Klasyfikacja gramatyczna leksemów polskich". *Język Polski* LIV(1): 3–13.

———. (1976a) *Cechy składniowe polskiego czasownika*. Wrocław: Ossolineum.

———. (1976b) „Kategoria rodzaju we współczesnym języku polskim". *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*. Wrocław: Ossolineum, 41–75.

———. (1977) „Kategorie gramatyczne liczebników we współczesnym języku polskim". *Studia Gramatyczne* I: 145–173.

———. (1996) „Homonimia a hasła w słownikach polskich". *Język Polski* LXXVI(4–5): 303–314.

Szober, Stanisław. (1920) „Sposoby łączenia liczebników zbiorowych z rzeczownikami". *Język Polski* V(1): 27–28.

———. (1922) „Sposoby łączenia liczebników głównych z rzeczownikami". *Język Polski* VII(5): 128–134.

———. (1928) „*Trzy piękne córki było nas u matki*: Formy podmiotu i orzeczenia w zdaniach z podmiotem logicznym, określonym przydawką liczebnikową". *Język Polski* XIII(4): 97–106.

Świdziński, Marek. (2001) „Limeryki koniugacyjne". Gruszczyński i in. (2001), 269–274.

Wiśniewski, Marek. (1990) „Czy liczebnik może być uznany za nadrzędnik dystrybucyjny szeregu rzeczownikowego?". *Studia Gramatyczne* IX: 87–97.

Wojdak, Piotr. (2000) „Z pogranicza odmienności i nieodmienności. pojęcie niezmorfologizowanej kategorii fleksyjnej i leksemu synkretycznego". *Polonica* XX: 233–251.

Woliński, Marcin. (2001) „Rodzajów w polszczyźnie jest osiem". Gruszczyński i in. (2001), 303–305.

———. (2002) „System znaczników morfosyntaktycznych w korpusie IPI PAN". Złożone do *Poloników*.

SUMMARY

Syntactic aspects of the morphosyntactic annotation in the IPI PAN corpus

The aim of this article is to discuss some *syntactic* factors influencing the *morphosyntactic* (part of speech and morphological) annotation of a large corpus of Polish designed and constructed at the Institute of Computer Science, Polish Academy of Sciences.

In particular, we argue that, in order to design a maximally transparent system of morphosyntactic annotation, one should be aware of and take into account such difficult areas of Polish syntax as the internal and external syntax of numeral phrases, the distribution of anaphoric forms and the rules governing the so-called negative concord.

We also discuss the impact of the syntactic context on the process of morphosyntactic disambiguation, showing that the rules of disambiguation between different forms of a single (f)lexeme must be based on a number of syntactic decisions, as when determining the case of a numeral in the subject position. Finally, we propose a criterion for disambiguating between the forms of different (f)lexemes and discuss some difficult cases of such inter-(f)lexemic disambiguation.