1

Analiza fleksyjna

Polszczyzna należy do języków o bogatej odmianie wyrazów[1]. Słownik języka polskiego notujący 200 000 haseł można uznać za obszerny. Hasłom takiego słownika odpowiada jednak kilka milionów różnych realizacji pojawiających się w tekstach. Co więcej, polski jest językiem fleksyjnym, co oznacza, że końcówki fleksyjne, stanowiące podstawowy środek tworzenia form odmiany, kumulują kilka funkcji gramatycznych. Sprawia to, że tworzenie form jest dużo mniej regularne niż w innych językach syntetycznych (językach aglutynacyjnych), gdzie każdy element doklejany do tworzonej formy pełni jedną funkcję gramatyczną. Dlatego pierwszą przeszkodą, którą trzeba pokonać przy komputerowej analizie tekstu polskiego, jest wykonanie analizy fleksyjnej. Jej celem jest powiązanie pewnych ciągów znaków tekstu wejściowego z abstrakcyjnymi jednostkami językowymi opisywanymi w słownikach języka polskiego. W największym uproszczeniu chodzi o rozpoznanie, że na przykład pojawiające się w tekście ciągi *psem* i *psach* odnoszą się do tej samej jednostki językowej – leksemu rzeczownikowego PIES, a ciąg *psim* jest tekstową realizacją innej jednostki – przymiotnika PSI. Wymaga to określenia zasad wyróżniania poszczególnych jednostek i klas jednostek.

Jednocześnie zadaniem opisu fleksyjnego jest ustalenie takiego repertuaru kategorii gramatycznych i ich wartości, aby przypisując je poszczególnym wyrazom występującym w tekście, można było odpowiednio precyzyjnie opisać ich funkcję składniową (ich uwarunkowania składniowe). Tak więc na przykład uznaje się, że formy rzeczownikowe są klasyfikowane m.in. ze względu na przypadek gramatyczny, aby dało się scharakteryzować różnicę gramatyczną między formą *psem* (narzędnik) i *psu* (celownik), przekładającą się na różne dopuszczalne konteksty użycia tych form. Dobranie odpowiednich kategorii gramatycznych i zbiorów ich wartości pozwala skatalogować wszystkie możliwe realizacje tekstowe form, a więc zestawić słownik fleksyjny.

Przedstawione tu podejście do fleksji jest powierzchowne. Jego celem jest tylko opracowanie klasyfikacji form, która nie musi objaśniać mechanizmów ich tworzenia (wyróżniania morfemów, opisu ich interakcji, uwarunkowań ortograficznych itd.).

---

[1] Określeń „wyraz" i „konstrukcja" będę używać nieformalnie, gdy bardziej chodzi o odwołanie do intuicji niż rygorystycznie zdefiniowanego pojęcia.

W tym rozdziale zostanie przedstawiona stosowana w niniejszej pracy koncepcja opisu fleksyjnego polszczyzny, która została zaimplementowana w analizatorze i generatorze fleksyjnym Morfeusz 2 (por. p. 1.3). Była ona dopracowywana stopniowo. Jej zrąb powstał na potrzeby znakowania Korpusu IPI PAN (Woliński i Przepiórkowski 2001; Woliński 2003; Przepiórkowski 2003b), a dalsza ewolucja była związana z analizatorem Morfeusz 2 (Woliński 2014). Koncepcja ta oparta jest przede wszystkim na pracach zespołu skupionego wokół Zygmunta Saloniego, których zwieńczeniem jest opracowanie *Słownika gramatycznego języka polskiego* (Saloni *et al.* 2007b, 2012, 2015, dalej: SGJP)[2].

Opracowanie systemu znakowania fleksyjnego, który byłby użyteczny w zastosowaniach komputerowych, wymaga podjęcia szeregu decyzji dotyczących stopnia szczegółowości opisu i poziomu wierności językoznawczej. Opis stworzony 10 lat temu dawał preferencję językoznawczej elegancji (np. kwestia rodzaju gramatycznego i sposób opisu form czasu przeszłego, zob. dalej). Obecnie prezentowana wersja zawiera kompromisy zwiększające jego komputerową użyteczność. Postanowiono nie wprowadzać zbyt szeroko rozróżnień, które dotyczą tylko niewielkich grup form lub tylko niewielu leksemów (konsekwencje tego założenia widać na przykład w opisie deprecjatywności, liczebników zbiorowych, form poprzyimkowych zaimków osobowych).

Przedstawiana koncepcja powstała w celu wykorzystania co najmniej na dwa sposoby: jako system znakowania korpusu, który byłby użyteczny dla jego użytkowników poprzez zapewnienie odpowiedniej szczegółowości wyszukiwania form fleksyjnych, oraz jako dane używane w dalszym przetwarzaniu, przede wszystkim w analizie składniowej.

1.1. Podstawowe pojęcia

Jednym z celów analizy fleksyjnej jest pogrupowanie wyrazów występujących w tekście w abstrakcyjne jednostki języka. Jednostki te nazywa się *leksemami* (wyrazami słownikowymi, por. Saloni 1974a). Pojęcie to nie jest absolutne, lecz związane z poziomem szczegółowości opisu przyjętego w danym słowniku. W słowniku notującym znaczenia za osobne leksemy mogą być uważane jednostki różniące się znaczeniem. W SGJP przyjęto, że kryterium wyróżnienia leksemów są tylko różnice własności fleksyjnych, a pomocniczo – własności składniowych. Leksem jest jednostką abstrakcyjną, tak więc leksemy nie występują w tekstach; są wynikiem interpretacji jako elementy wykoncypowanego systemu językowego. Z każdym leksemem jest związany identyfikator, nazywany *lematem* (por. p. 1.9).

[2] Prezentacja w tym rozdziale wykorzystuje wcześniejsze prace autora, w szczególności artykuł Woliński (2003). Zawiera jednak niepublikowane wcześniej rozstrzygnięcia wprowadzone ostatnio, w szczególności dotyczące reprezentacji rodzaju gramatycznego, zob. p. 1.6.3.

Zapis tekstów przyjęty dla języka polskiego pozwala w naturalny sposób wyróżnić *słowa*: najdłuższe ciągi znaków niezawierające odstępów ani znaków interpunkcyjnych (co nie znaczy, że żadnych znaków nieliterowych). Leksemy najczęściej przejawiają się w tekstach w postaci słów, na przykład słowo *chomikiem* można uznać za realizację leksemu CHOMIK. Jednak niekiedy przedmiotem interpretacji fleksyjnej powinny być fragmenty tak wyróżnionych słów. Wyrazistym przykładem jest słowo *gdybyście* występujące w wypowiedzeniu:

(1) *Wiedzielibyście, gdybyście słuchali.*

Próba interpretacji tego słowa w całości prowadziłaby zapewne do uznania go za spójnik podrzędny, jako że wprowadza ono zdanie podrzędne *gdybyście słuchali*. Jednocześnie widać uwarunkowanie osobowo-liczbowe: słowo *gdybyście* może wystąpić razem ze *słuchali*, ale nie ze *słuchał* ani *słuchaliśmy*. Fakty te sugerują wprowadzenie bytu bardzo nietypowego: spójnika odmiennego przez osobę i liczbę (por. Świdziński 1981). Jednak prostszym rozwiązaniem jest uznanie, że osobnej interpretacji powinny podlegać fragmenty tego słowa *gdyby* i *ście*. Element podlegający interpretacji fleksyjnej będzie dalej nazywany *segmentem* (w terminologii angielskiej w takim znaczeniu jest zwykle używany wyraz *token*). Tak więc ciąg znaków *chomikiem* jest jednocześnie słowem i segmentem, a ciąg *gdybyście* jest słowem składającym się z segmentów *gdyby* i *ście* (zagadnienie wyróżniania segmentów rozwinięto w p. 1.5).

Segmenty są niezinterpretowanymi ciągami znaków, leksemy są abstrakcyjnymi jednostkami języka. Bytem, który wiąże te dwa poziomy, są formy fleksyjne. *Forma fleksyjna* to interpretacja segmentu przypisująca go do konkretnego leksemu i opisująca jego własności gramatyczne poprzez określenie wartości pewnych kategorii gramatycznych (por. p. 1.6). Na przykład segment *chomikiem* może zostać zinterpretowany jako forma fleksyjna rzeczownikowego leksemu CHOMIK o wartości narzędnika kategorii przypadka i pojedynczej wartości liczby.

Inaczej mówiąc, segmenty są tekstowymi wykładnikami (realizacjami) form fleksyjnych (stosując skrót myślowy, są wykładnikami leksemów).

Technicznie formę fleksyjną można uznać za trójkę uporządkowaną ⟨*segment, lemat, znacznik fleksyjny*⟩, gdzie *znacznik fleksyjny* (ang. *tag*) stanowi zwarty zapis cech gramatycznych form (zob. p. 1.10). Zbiór wszystkich form fleksyjnych danego leksemu nazywa się jego *paradygmatem (fleksyjnym)*.

Analiza fleksyjna[3] polega na wskazaniu dla danego segmentu wszystkich form wszystkich leksemów, których może on być wykładnikiem. W procesie

[3] W środowisku komputerowego przetwarzania języka bywa też równoważnie używany termin *analiza morfologiczna*. Jednak w językoznawstwie termin ten obejmuje badanie budowy wewnętrznej wyrazów, w szczególności określanie ich elementarnych składników – morfemów, co nie ma miejsca w przedstawionym tu opisie. Ponadto niektórzy badacze do analizy morfologicznej (w odróżnieniu od fleksyjnej) włączają też opis derywacji.

We wcześniejszych pracach, np. Woliński i Przepiórkowski 2001, stosowane jest określenie *analiza morfosyntaktyczna* dla podkreślenia, że część podawanych cech ma charakter

Tabela 1.1. Przykład wyników analizy fleksyjnej tekstu *Mam próbkę analizy fleksyjnej.* wykonanej przez program Morfeusz 2 SGJP

		segment	lemat	znacznik fleksyjny
0	1	*Mam*	MAMA MAMIĆ MIEĆ	subst:pl:gen:f impt:sg:sec:imperf fin:sg:pri:imperf
1	2	*próbkę*	PRÓBKA	subst:sg:acc:f
2	3	*analizy*	ANALIZA	subst:sg:gen:f subst:pl:nom.acc.voc:f
3	4	*fleksyjnej*	FLEKSYJNY	adj:sg:gen.dat.loc:f:pos
4	5	.	.	interp

tym nie uwzględnia się kontekstu, w którym wystąpił dany segment, wyniki są więc często niejednoznaczne.

Ujednoznacznianiem fleksyjnym nazywa się określanie na podstawie kontekstu, jaką formę fleksyjną realizuje dane wystąpienie segmentu. Następujące po sobie analizę i ujednoznacznianie fleksyjne nazywa się *tagowaniem* (ang. *tagging*).

Celem *hasłowania* (*lematyzacji*) jest wskazanie dla danego segmentu leksemu, którego formy jest on wykładnikiem. Jest to więc tagowanie ograniczone tylko do części informacji o formach – do lematów. Przybliżone hasłowanie polegające na odcięciu od słów części zmieniającej się przy odmianie bywa nazywane *stemowaniem* (ang. *stemming*). Metoda ta ma sens w odniesieniu do języków o ograniczonej fleksji, ale dla języka polskiego daje wyniki wysoce niezadowalające.

Operacją odwrotną do analizy fleksyjnej jest *synteza fleksyjna* – utworzenie wszystkich form odmiany leksemu na podstawie jego lematu.

W celu zilustrowania wymienionych pojęć w tabeli 1.1 przedstawiono przykład wyników analizy fleksyjnej. Każdy wiersz tabeli zawiera jedną formę fleksyjną, kreski oddzielają grupy interpretacji dla poszczególnych segmentów. Segmentowi *mam* zostały przypisane trzy interpretacje: jako forma liczby mnogiej rzeczownika MAMA, jako forma trybu rozkazującego czasownika MAMIĆ i wreszcie jako forma czasu teraźniejszego czasownika MIEĆ. Segment *analizy* został jednoznacznie przypisany do lematu ANALIZA, może on jednak być interpretowany zarówno jako forma liczby pojedynczej, jak i mnogiej – w różnych przypadkach.

---

składniowy, a nie czysto fleksyjny. Dotyczy to na przykład podawania rodzaju dla rzeczowników, skoro nie odmieniają się one przez rodzaj. Określenie to jest mylące, bo sugeruje analizę, która obejmuje morfologię i składnię.

Dlatego w niniejszej pracy postanowiłem pozostać przy odrobinę za wąskim terminie *analiza fleksyjna*.

Wielość interpretacji dla jednego segmentu jest w języku polskim zjawiskiem bardzo częstym, warto więc wprowadzić dwa pojęcia opisujące takie sytuacje (por. Świdziński *et al.* 2002). *Homonimia* to równość wykładników form należących do różnych leksemów. *Synkretyzm* to równość wykładników różnych form należących do tego samego leksemu[4]. W przytoczonym przykładzie segment *Mam* może być wykładnikiem homonimicznych form trzech różnych leksemów. Synkretyczne są różne formy leksemu ANALIZA o wykładniku *analizy*.

Stosowane znaczniki fleksyjne są pozycyjne. Pierwsza pozycja określa klasę gramatyczną (część mowy), następne reprezentują wartości kategorii gramatycznych przysługujących danej klasie. Na przykład znacznik subst oznacza rzeczownik, a po nim następują wartości liczby, przypadka i rodzaju. Oznaczenia są w większości skrótami łacińskich nazw wartości. Konstrukcja przyjętych kategorii gramatycznych i dopuszczalne wartości poszczególnych kategorii zostały omówione w dalszych podrozdziałach.

1.2. SGJP

Prezentowany tu system analizy fleksyjnej stanowi przystosowaną do potrzeb analizy komputerowej postać opisu fleksji w *Słowniku gramatycznym języka polskiego* (Saloni *et al.* 2007b, 2012, 2015).

Ideę opracowania słownika pokazującego odmianę polskich wyrazów i cechy gramatyczne ich form powziął Zygmunt Saloni pod wpływem analogicznego słownika języka rosyjskiego (Zalizniak 1977). Realizacja tego zamierzenia trwała około 30 lat, w czasie których Saloni wraz z nieformalnym zespołem, obejmującym pracowników kilku placówek naukowych, zajmował się opracowaniem poszczególnych części materiału.

Prace rozpoczęły się od analizy opisów gramatycznych największego dostępnego wówczas słownika języka polskiego (Doroszewski 1958–1969). Informacja zawarta w tych opisach pozwalała tylko na przybliżone ustalenie sposobu odmiany, była jednak ważnym punktem wyjścia. W końcu lat 70. prace były prowadzone na papierowych fiszkach (ok. 130 000 sztuk, por. Saloni *et al.* 2007a). Użytecznym materiałem był także indeks *a tergo* do słownika Doroszewskiego (Grzegorczykowa i Puzynina 1973). Indeks ten został później zdigitalizowany przez Roberta Wołosza.

Istotnym osiągnięciem było opracowanie przez Włodzimierza Gruszczyńskiego opisu odmiany (deklinacji) polskich rzeczowników pospolitych (Gruszczyński 1989).

---

[4] Niektórzy badacze obejmują pojęciem homonimii oba wymienione zjawiska, wyróżniając homonimię międzyparadygmatyczną (=homonimia) i wewnątrzparadygmatyczną (=synkretyzm).

Pierwszym efektem prac grupy obejmującym materiał fleksyjny jako całość był przygotowany do publikacji przez Saloniego schematyczny indeks form fleksyjnych zainicjowany przez Jana Tokarskiego (Tokarski 1993). W dziele tym zakończenia form odmiany zostały powiązane z zakończeniami form podstawowych. System fleksyjny potraktowano jako potencję, a więc jako opis mechanizmów, a nie odmiany konkretnych leksemów. Indeks wykorzystano do konstrukcji kilku programów komputerowych (Szafran 1993; Wołosz 2005; Woliński 2006b). Indeks był opracowywany w postaci elektronicznej – pliku tekstowego z rygorystycznie zadanymi kolumnami reprezentującymi poszczególne informacje[5].

Saloni podjął prace nad opisem odmiany czasowników (koniugacji) na podstawie wcześniejszych prac Tokarskiego, który był również autorem systemu wzorców koniugacyjnych używanego w słowniku Doroszewskiego (Tokarski 1951). Doprowadziły one do wydania precyzyjnego słownika odmiany czasowników (Saloni 2001, 2007). Na tym etapie z grupą zaczął współpracować Marcin Woliński, który stopniowo przejął zadanie formowania komputerowego modelu opracowywanych danych. Tak więc *Czasownik polski* od pierwszego wydania był publikowany na podstawie relacyjnej bazy danych (Saloni i Woliński 2004, 2003). Dane te zostały także wykorzystane od razu do komputerowej analizy fleksyjnej.

Na tym etapie realne stało się myślenie o wydaniu słownika fleksyjnego obejmującego cały materiał języka polskiego (Gruszczyński i Saloni 2006; Gruszczyński 2001; Saloni i Woliński 2005). W ramach podjętych prac opis wszystkich części mowy został stopniowo doprowadzony do stopnia precyzji reprezentowanego przez *Czasownik polski*. Dane opisujące poszczególne części mowy zostały w szczególności sprowadzone do wspólnego modelu komputerowego (Woliński 2009).

Odmiana poszczególnych leksemów jest w SGJP opisywana poprzez przypisanie wzoru fleksyjnego. Podstawowa zasada tworzenia wzorów fleksyjnych polega na odcięciu od wszystkich form odmiany wspólnej części początkowej i potraktowaniu dalszych części zmiennych jako wzoru fleksyjnego. Wzory tworzone według takiej mechanicznej zasady są dość liczne, obecna wersja słownika operuje 1116 wzorami.

Początkowo planowane było wydanie słownika zarówno w postaci papierowej, jak i elektronicznej, jednak po dyskusji uznano, że celowe będzie ograniczenie projektu do wersji elektronicznej. Tak więc dwa pierwsze wydania (Saloni *et al.* 2007b, 2012) miały postać programu komputerowego dystrybuowanego na płycie CD i towarzyszącej książki ze wstępem teoretycznym (Saloni *et al.* 2007a). Trzecie wydanie przyjęło postać aplikacji internetowej http://sgjp.pl (Saloni *et al.* 2015; Woliński i Kieraś 2016). Na potrzeby tego wydania wzory fleksyjne poddano rewizji i opracowano ich nową klasyfikację (Saloni 2016).

[5] Plik ten jest dostępny do pobrania pod adresem http://sgjp.pl/siat/.

W trzecim wydaniu SGJP osiągnął wielkość ponad 330 000 leksemów[6] odpowiadających prawie 4 300 000 wykładników tekstowych. Można przyjąć, że słownik obejmuje cały materiał leksykalny słownika Doroszewskiego rozszerzony o formy ekscerpowane ze współczesnych korpusów i innych słowników. Wypada jednak zaznaczyć, że słownik nie zawiera informacji frekwencyjnej i nie był uzupełniany na podstawie frekwencji wyrazów w korpusie. Część siatki haseł stanowią regularnie derywowane leksemy pochodne, niektóre z nich mniej lub bardziej potencjalne, np. w słowniku notowane są rzeczownikowe nazwy cech typu BIAŁOŚĆ (skoro jest BIAŁY) i WYŻLOŚĆ (skoro jest WYŻLI).

SGJP jest źródłem danych fleksyjnych dla *Wielkiego słownika języka polskiego* (Żmigrodzki 2013–2018), a także jest często przywoływany przez *Wikisłownik* (https://pl.wiktionary.org/).

1.3. Morfeusz SGJP

Pierwsza wersja analizatora fleksyjnego Morfeusz została opracowana około roku 2001. Początkowo dane programu stanowił indeks Tokarskiego (1993, SIaT) skonfrontowany z listą haseł słownika Doroszewskiego (Doroszewski 1958–1969). Opis czasowników pochodził z *Czasownika polskiego* (Saloni 2001). Ta wersja programu została retrospektywnie nazwana *Morfeuszem SIaT* (Woliński 2006b).

Następnie te przybliżone dane były zastępowane bardziej precyzyjnymi danymi SGJP – wynik został udostępniony jako *Morfeusz SGJP*. W kolejnym etapie program zaimplementowano od nowa w celu zwiększenia funkcjonalności (Woliński 2014) – powstała wersja *Morfeusz 2*. Do modułu analizy dodano moduł syntezy. Informacje dostarczane przez program wzbogacono o kwalifikatory ze słownika i prostą klasyfikację nazw własnych. Ważnym nowym elementem jest możliwość zadawania reguł segmentacji tekstu, co umożliwiło skonstruowanie analizatora tekstów historycznych uwzględniającego zmieniające się reguły pisowni łącznej i rozdzielnej (por. p. 1.5).

Podjęto także prace nad połączeniem SGJP z danymi tworzonego społecznościowo słownika fleksyjnego SJP.pl. W ich wyniku powstał słownik Polimorf (Woliński *et al.* 2012). Obecnie program jest dystrybuowany z oboma wariantami słownika (tabela 1.2).

Morfeusz stanowi wyłącznie interfejs do danych słownikowych – nie zawiera modułu interpretującego słowa spoza słownika ani mechanizmu ujednoznaczniającego wyniki. Ten ostatni fakt jest istotny, ponieważ w języku

[6] Wypada zaznaczyć, że leksemy w SGJP są wyodrębnione inaczej niż hasła słownika Doroszewskiego i inaczej niż leksemy analizatora Morfeusz (por. tabela 1.2 i p. 1.7.1). W szczególności w SGJP osobne hasła stanowią odsłowniki i imiesłowy przymiotnikowe, które w słowniku Doroszewskiego i w analizatorze Morfeusz stanowią część leksemu czasownikowego.

Tabela 1.2. Słowniki dystrybuowane z programem *Morfeusz 2* (dane liczbowe dla wersji z końca 2017 roku)

słownik	leksemy	wykładniki
SGJP	264 166	4 037 250
Polimorf	315 055	3 844 535

polskim bardzo częstym zjawiskiem jest homonimia i synkretyzm. Dokładne określenie liczbowe ich poziomu istotnie zależy od przyjętego sposobu znakowania tekstu. Jeżeli przyjąć opisywany tu system zaimplementowany w analizatorze Morfeusz, okazuje się, że 34,5% segmentów w tekście może być interpretowanych jako forma więcej niż jednego leksemu, natomiast 68,5% segmentów ujawnia jakąkolwiek formę niejednoznaczności, w tym dotyczącą wartości przypisanych kategorii gramatycznych (obliczenia wykonano na zrównoważonym 300-milionowym NKJP).

Warto zwrócić uwagę, że w niektórych pracach bardziej użyteczne są wyniki niejednoznaczne niż interpretacje ujednoznacznione statystycznie. Analizę fleksyjną można stosunkowo łatwo wykonać automatycznie z dużą dokładnością. Jest to jedynie kwestia zgromadzenia odpowiednio bogatego słownika fleksyjnego. Ujednoznacznianie interpretacji konkretnych wystąpień wyrazów wymaga odwoływania się do zjawisk składniowych i semantycznych, przez co może być wykonane algorytmicznie tylko w przybliżeniu. Najlepsze statystyczne programy ujednoznaczniające (tagery) dla języka polskiego osiągają skuteczność ok. 93% (Kobyliński i Kieraś 2016). To oznacza, że, przyjmując niezależność ujednoznacznienia na poszczególnych pozycjach tekstu, można oszacować prawdopodobieństwo poprawnego zinterpretowania całego zdania zawierającego 15 wyrazów na mniej niż ½. Niestety wykluczenie poprawnej interpretacji choćby jednego wyrazu może spowodować błędny wynik algorytmu interpretującego zdanie jako całość, np. podczas analizy składniowej. Jednak w automatycznej analizie składniowej wielość interpretacji fleksyjnych nie musi przeszkadzać – to właśnie wykonanie analizy składniowej wypowiedzenia może dać ujednoznacznienie na poziomie fleksyjnym.

Można oczekiwać, że prawdopodobieństwo błędu ujednoznacznienia jest większe dla form rzadkich. Ale to te właśnie formy mogą być interesujące dla użytkownika korpusu. Dlatego przy konstrukcji Narodowego Korpusu Języka Polskiego zdecydowano, że dostępne dla użytkownika muszą być zarówno wszystkie interpretacje fleksyjne przypisane bezkontekstowo, jak i wybrana spośród nich interpretacja właściwa ze względu na kontekst. W konsekwencji wyszukanie interpretacji nietypowych może wymagać przejrzenia większej liczby potencjalnych kontekstów, ale nie jest już niemożliwe.

Morfeusz wewnętrznie używa minimalnych automatów skończonych do zwartej reprezentacji słowników (Woliński 2006b). Program jest udostępniany w wersji gotowej do użycia na główne platformy sprzętowe (Linux, OS X, Win-

dows). Ma on postać biblioteki dynamicznej (API C++), jako że podstawowym odbiorcą ma być programista wbudowujący analizator w tworzone narzędzia przetwarzania tekstu. Udostępniane są także moduły pozwalające na użycie Morfeusza w programach w Pythonie, Perlu, Javie, SWI Prologu oraz interfejs graficzny dla mniej technicznie zaawansowanych użytkowników.

Przyjęty system znakowania tekstu (ang. *tagset*) jest głównym przedmiotem rozważań w pozostałej części tego rozdziału (por. także Woliński 2003; Przepiórkowski i Woliński 2003a,b,c; Przepiórkowski 2009; Woliński 2001). System ten został pierwotnie opracowany na potrzeby znakowania Korpusu IPI PAN (Przepiórkowski *et al.* 2003; Przepiórkowski 2004a; Woliński i Przepiórkowski 2001), a następnie jego wariant został użyty do znakowania Narodowego Korpusu Języka Polskiego (NKJP, Przepiórkowski *et al.* 2012). System znaczników Morfeusza nie był nigdy dokładnie tożsamy z tagsetem NKJP. Przedstawiona w tej pracy najnowsza wersja zmniejsza rozbieżności między tymi systemami.

Istotną kwestią dla użyteczności analizatora fleksyjnego jest licencja, na jakiej jest on udostępniony. Analizator taki działa poprzez kopiowanie do swoich wyników fragmentów używanego słownika fleksyjnego. Dlatego wynik analizy staje się utworem zależnym w stosunku do użytego słownika fleksyjnego, w związku z czym ograniczenia licencyjne słownika zaczynają dotyczyć wyników analizatora[7]. W związku z powyższym zdecydowano, żeby program *Morfeusz 2* i towarzyszące mu słowniki fleksyjne SGJP i Polimorf były dystrybuowane na bardzo liberalnej licencji BSD, która dopuszcza dowolne zastosowania, żądając jedynie zachowania informacji o autorstwie.

Morfeusz jest dość powszechnie używany przez polskie środowisko naukowe. Program został użyty do znakowania Korpusu IPI PAN (Przepiórkowski 2004a) i Narodowego Korpusu Języka Polskiego (Przepiórkowski *et al.* 2012). Na jego bazie opracowano liczne programy ujednoznaczniające analizy (tagery): tager HMM (Dębowski 2004), TaKIPI (Piasecki 2007), PANTERA (Acedański 2010), WMBT (Radziszewski i Śniatowski 2011), Concraft-pl (Waszczuk 2012), WCRFT (Radziszewski 2013), PoliTa (Kobyliński 2014), Toygger (Krasnowska-Kieraś 2017). Morfeusz został zintegrowany z systemami tworzenia płytkich parserów SproUT (Piskorski *et al.* 2004) i Spejd (Przepiórkowski 2008), a także z narzędziem opisu odmiany jednostek wieloczłonowych Multiflex (Savary 2005) i Toposław (Marciniak *et al.* 2011). Wreszcie Morfeusz jest składnikiem opisywanego tu parsera Świgra 2 i parsera POLFIE (Patejuk 2015).

1.4. Dystrybucyjne podejście do fleksji

Niniejsza praca dotyczy przetwarzania języka w jego odmianie pisanej. Obiektywnie istniejącym przedmiotem badań są więc teksty stanowiące cią-

[7] Taką opinię uzyskano od prawnika specjalizującego się w prawie autorskim.