Elżbieta Awramiuk
Urszula Andrejewicz
(Białystok)

Problemy segmentacyjne w analizie morfologicznej

Segmentacja tekstu jest podstawowym zabiegiem wykonywanym przy pracach językoznawczych i leksykograficznych. Decyzje segmentacyjne wywierają wpływ na wyniki statystyczne[1], na obraz polskiej leksyki zawarty w słownikach. Przy obecnym stopniu automatyzacji prac nad korpusami tekstów pierwszej segmentacji tekstu pisanego na słowa, automatycznie – od spacji do spacji – dokonuje komputer[2]. Analiza morfologiczna tekstu pisanego, rozumiana jako przypisanie pojawiającym się w tekście słowom charakterystyki znaczeniowej i gramatycznej oraz przyporządkowanie wyodrębnionych w ten sposób form wyrazowych do odpowiednich leksemów[3], może być problematyczna, gdy granice słów nie odpowiadają formom wyrazowym. W takim wypadku mechaniczne wyodrębnienie słów od spacji do spacji nie wystarcza. Celem niniejszego artykułu jest opis form wyrazowych nieodpowiadających pojedynczym słowom oraz uporządkowanie dotyczących ich informacji gramatycznych w taki sposób, aby wyeksponować regularność tego zjawiska.

Ustalenia terminologiczne

W niniejszym tekście słowo uważa się za unilateralną jednostkę językową, rozumianą jako ciąg liter od spacji do spacji (obserwacje dotyczą języka pisanego). Słowo rozumiane jest tu zatem jako jednostka czysto tekstowa, reprezentująca jedynie kształt (postać graficzną). Znak językowy (w rozumieniu desaussurowskim) jest parą uporządkowaną (a, x), gdzie a oznacza kształt (formę), a x – znaczenie (sens). Nadanie słowu sensu (x) przenosi je na poziom bilateralny: tak zinterpretowane słowo staje się formą wyrazową. Formy wyrazowe, których wykładnikiem tekstowym jest to samo słowo (a), ale z przyporządkowanymi dwoma znaczeniami (x_1) oraz (x_2), nazywa się homonimami[4].

---

1 Por. I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, *Słownik frekwencyjny współczesnej polszczyzny*, Kraków 1990 (dalej: SFWP); E. Awramiuk E., *Wpływ odstępstw od segmentacji ortograficznej na wyniki statystyczne "Słownika frekwencyjnego polszczyzny współczesnej"*, „Roczniki Humanistyczne" XLIX-L (2001-2002), z. 6, s. 31-43.

2 O fleksji w analizie maszynowej pisał już w latach 70. Jan Tokarski, por. tegoż *Fleksja polska*, Warszawa 1978, s. 9-11.

3 W pracach leksykograficznych oznacza to przyporządkowanie formom wyrazowym form hasłowych.

4 Por. Z. Saloni, *Homonimia a hasła w słownikach polskich*, „Język Polski" LXXVI (1996), z. 4-5, s. 303-314.

Proces wyróżniania jednostek leksykalnych podczas automatycznej analizy tekstu obejmuje etap segmentacji i bilateralizacji tekstu, por. analizę zdania:

/1/ *Jan śpiewał.*

I. Segmentacja tekstu i wyróżnienie jednostek unilateralnych (słów):
/1a/ *Jan / śpiewał.*

II. Bilateralizacja wyróżnionych jednostek i wyodrębnienie form wyrazowych:
/1b/ *Jan* *śpiewał*
M. lp., r.m. JAN (S) 3 os., lp., cz. przesz. ŚPIEWAĆ (V)

Dokładniej, etap bilateralizacji dzieli się na dwie części, co można zaobserwować na przykładzie wypowiedzenia /2/, por.:

/2/ *Jan będzie śpiewał.*

I. Segmentacja tekstu i wyróżnienie jednostek unilateralnych (słów):
/2a/ *Jan / będzie / śpiewał.*

IIa. Bilateralizacja wyróżnionych jednostek i wyodrębnienie wyrazów morfologicznych[5]:

/2b/ *Jan*	*będzie*	*śpiewał*
M. lp., r.m. JAN (S)	3 os., lp., cz. przysz. BYĆ (V)	3 os., lp., cz. przesz. ŚPIEWAĆ (V)

IIb. Wyodrębnienie form wyrazowych:
/2c/ *Jan* *będzie śpiewał*
M. lp., r.m. JAN (*S*) 3 os., lp., cz. przysz. ŚPIEWAĆ (*V*)

W szczególnych wypadkach poddane analizie morfologicznej słowo nie jest bowiem wykładnikiem formy wyrazowej. Może ono stanowić element składowy formy złożonej, jak w przykładach /2/ i /3/, wykładnik dwóch różnych form wyrazowych, jak w przykładzie /4/, lub jeden ze składników nieciągłej jednostki leksykalnej, jak w przykładzie /5/:

/3/ *Będzie pisał z wielkim zapałem.*
/4/ *Spojrzał nań z lekką niechęcią.*
/5/ *Śmiał się głośno i serdecznie.*

Przy ujęciu unilateralnym zdania /3/-/5/ zbudowane są z pięciu słów. Podczas interpretacji poszczególnych znaków i dopisywania do nich konkretnego znaczenia (określania ich funkcji gramatycznych), można zauważyć, że w przykładzie /3/ słowo *będzie* wraz ze słowem *pisał* stanowi formę gramatyczną czasownika PISAĆ (3 os. lp. czas

[5] Por. J. Bień, Z. Saloni, *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*, „Prace Filologiczne" XXXI (1982), s. 31-45.

przyszły)[6], w przykładzie /4/ słowo *nań* odsyła do dwóch form wyrazowych, reprezentujących dwa leksemy: przyimek NA i zaimek ON, z kolei dwa słowa *śmiał się* z przykładu /5/ razem stanowią nieciągłą formę wyrazową leksemu ŚMIAĆ SIĘ.

Przedmiotem zainteresowania w artykule uczyniono jednostki, które podczas analizy morfologicznej okazują się elementem jednostki nielinearnej. Chodzi o sytuacje niepokrywania się granic jednostek leksykalnych ustalanych w procesie bilateralizacji z formalnym podziałem uzyskiwanym w wyniku automatycznej segmentacji od spacji do spacji, jak w przykładach /3/-/5/. W krąg obserwacji wchodzi również zjawisko homonimii, zarówno międzyparadygmatycznej, czyli sytuacji, w których słowo stanowi wykładnik form dwóch leksemów (np. *damy* jako forma czasownika DAĆ i rzeczownika DAMA), jak i wewnątrzparadygmatycznej, czyli homonimii form wyrazowych jednego leksemu (np. *kobiety* jako D. lp. lub M. l.mn. rzeczownika KOBIETA), w takim zakresie, w jakim wiąże się z nimi problem nielinearności.

Aby ustalić, jakie elementy mogą stanowić składnik ciągu słów reprezentujących jedną formę gramatyczną, należy określić paradygmaty fleksyjne dla poszczególnych części mowy. Wiąże się to ściśle z klasyfikacją leksemów. Podział na części mowy przyjęto za Zygmuntem Salonim[7] ze względu na dyscyplinę formalną i przyjęte tam kryterium morfologiczne.

Analizie zostaną poddane następujące sytuacje niepokrywania się ortograficznej granicy słowa z granicą jednostki leksykalnej:

1) słowo jako jeden z elementów analitycznej formy wyrazowej;
2) słowo jako wykładnik dwóch form wyrazowych;
3) słowo jako składnik nieciągłej jednostki leksykalnej.

Słowo jako jeden z elementów analitycznej formy wyrazowej

Niektóre znaki językowe w wyniku bilateralizacji okazują się członami złożonych form fleksyjnych. Przyjęta klasyfikacja leksemów pozwala zakres tego typu jednostek ograniczyć do dwóch wypadków: form stopnia wyższego i najwyższego przymiotników oraz analitycznych form fleksyjnych czasowników.

Stopniowanie opisowe przymiotników jakościowych odbywa się poprzez zastosowanie słów *bardziej* i *najbardziej*, które w przyjętej klasyfikacji są formami przysłówkowymi przymiotnika, tak jak wszystkie regularnie tworzone przysłówki odprzymiotnikowe. Leksemy przymiotnikowe dzielą się bowiem na dwa podzbiory: podzbiór *A(adv)*, do którego należą formy przysłówkowe, oraz na podzbiór *A(-adv)*, obejmujący wszystkie formy pozostałe. Tradycyjnie wyróżniany przysłówek *bardzo* stanowi formę przysłówkową[8] leksemu przymiotnikowego WIELKI, supletywną wobec podstawy. Formy

[6] Określenie tej jednostki jako 3. os. czasu przeszłego odpowiada intuicji językowej użytkownika języka. W szkole czas przyszły złożony definiuje się poprzez regułę: czas przeszły BYĆ + 3 os. czasu przeszłego czasownika X = czas przyszły czasownika X. Schemat taki sprawdza się przy syntezie form (czyli tworzeniu należących do danego leksemu form wyrazowych o odpowiedniej charakterystyce gramatycznej), natomiast przy automatycznej analizie fleksyjnej nie, ponieważ tu funkcja form złożonych nie wynika z sumowania się funkcji ich składników, lecz jest pełniona przez nie obie wspólnie.

[7] Z. Saloni, *Klasyfikacja gramatyczna leksemów polskich*, „Język Polski" LIV (1974), s. 3-13, 93-101.

[8] Dla wygody pozostańmy przy określeniu "przysłówek" dla takich form wyrazowych, które należą do leksemu przymiotnikowego (czyli regularnie tworzonych przysłówków odprzymiotnikowych).

kategorii stopnia na przykład leksemu przymiotnikowego PONURY mają zatem następującą postać:

A(-adv): ponury, bardziej ponury, najbardziej ponury;
A(adv): ponuro, bardziej ponuro, najbardziej ponuro.

Każde wystąpienie słowa zinterpretowanego podczas bilateralizacji jako przymiotnik lub przysłówek odprzymiotnikowy wymaga dalszej analizy kontekstu i podjęcia decyzji, czy wyróżniony w jej przebiegu wyraz morfologiczny jest wykładnikiem syntetycznej formy wyrazowej czy elementu formy analitycznej. W procesie bilateralizacji wszystkich leksemów przymiotnikowych powstają jednostki homonimiczne – dokładnie w taki sam sposób:

(a, x_1): A (stopień równy) + Ø[9]
(a, x_2) : A (stopień równy) + [bardziej] = A (stopień wyższy)
(a, x_3) : A (stopień równy) + [najbardziej] = A (stopień najwyższy)

Czasownikowe struktury analityczne stanowią nie tylko jedność semantyczną, ale i funkcjonalną: to połączenie dwóch wyrazów morfologicznych, z których jedna jest wielofunkcyjna. Pod względem składniowym są one równoważne czasownikowym formom prostym. Taka struktura cechuje formy:

- czasu przyszłego złożonego[10] (z bezokolicznika[11] i formy wyrazowej na *-ł, -ła, -ło, -li* lub *-ły* czasowników niedokonanych i formy czasu przyszłego czasownika BYĆ)[12], np. *będę chodzić, będę chodziła*;
- czasu zaprzeszłego (złożone z formy czasu przeszłego czasownika i 3 os. czasu przeszłego BYĆ), np. *robił był*[13];
- analityczne trybu przypuszczającego w czasie nieprzeszłym[14] (złożone z partykuły BY z ewentualnie dołączonymi morfemami osobowymi i formy 3 os. czasu przeszłego trybu oznajmującego dowolnego czasownika), np. *bym pisał, by pisała*;
- analityczne trybu przypuszczającego w czasie przeszłym (złożone z formy prostej trybu przypuszczającego czasownika BYĆ i 3 os. czasu przeszłego dowolnego czasownika), np. *byłbym pisał*;

Leksykalne sposoby wyrażania różnic natężenia cechy (takie jak *trochę, całkiem*) pozostają poza obszarem naszej obserwacji.

9 Brak oczekiwanego elementu jest znaczący, gdyż pozwala zdefiniować analizowaną jednostkę przez przeciwstawienie jej innym.

10 Pomijamy tu tzw. czas przyszły potencjalny (np. *mam pisać*) i czas przeszły potencjalny (np. *miałem pisać*), które rzadko zalicza się do paradygmatu konkretnego leksemu czasownikowego (por. J. Tokarski 1978, s. 207).

11 Dokładniej: wyraz morfologiczny o kształcie bezokolicznika. Dalej będą używane skrótowo pojęcia: bezokolicznik, forma czasu przyszłego itp.

12 Por. polemikę H. Wróbla (1995) z R. Laskowskim (1998) na temat form czasownika BYĆ w funkcji flektywu (wykładnika funkcji gramatycznych, morfemu fleksyjnego w analitycznych formach czasownika).

13 Czas zaprzeszły jest strukturą o wyraźnym już dzisiaj nacechowaniu archaicznym. Jego prawdopodobieństwo pojawienia się w tekstach pisanych jest niewielkie, jednak istnieje. Zwiększa się ono oczywiście, jeśli analizie poddawane są teksty starsze lub współczesne, ale stylizowane na dawne.

14 Formy nieprzeszłe mogą odnosić się do przeszłości (*wczoraj by pisał*), teraźniejszości (*dzisiaj by pisał*) i przyszłości (*jutro by pisał*). Za ich pomocą wyrażany jest tzw. tryb potencjalny. Por. M. Bańko, *Wykłady z polskiej składni*, Warszawa 2002, s. 95-96.

– analityczne trybu rozkazującego (złożone z partykuły NIECH, NIECHAJ i form 3 os. czasu teraźniejszego czasowników niedokonanych oraz 3 os. czasu przyszłego prostego czasowników dokonanych), np. *niech śpiewa, niechaj śpiewa, niech zaśpiewa, niechaj zaśpiewa*[15].

Powyższy przegląd paradygmatu czasownika pozwala wyodrębnić zjawiska bardzo regularne. Obraz uwikłania czasownikowych form czasu przeszłego w homonimię (na przykładzie czasownika PISAĆ, z ograniczeniem do form rodzaju męskiego[16]) przedstawia się następująco:

(*a*, x_1): 1 os. lp., r. m., cz. przeszły + Ø (np. *pisałem*)
(*a*, x_2): 1 os. lp., r. m., cz. przeszły + [*był*] = 1 os. lp., r. m., cz. zaprzeszły

(*b*, x_1): 2 os. lp., r. m., cz. przeszły + Ø (np. *pisałeś*)
(*b*, x_2): 2 os. lp., r. m., cz. przeszły + [*był*] = 2 os. lp., r. m., cz. zaprzeszły

(*c*, x_1): 3 os. lp., r. m., cz. przeszły + Ø (np. *pisał*)
(*c*, x_2): 3 os. lp., r. m., cz. przeszły + [*będę*] = 1 os. lp., r. m., cz. przyszły
(*c*, x_3): 3 os. lp., r. m., cz. przeszły + [*będziesz*] = 2 os. lp., r. m., cz. przyszły
(*c*, x_4): 3 os. lp., r. m., cz. przeszły + [*będzie*] = 3 os. lp., r. m., cz. przyszły
(*c*, x_5): 3 os. lp., r. m., cz. przeszły + [*był*] = 3 os. lp., r. m., cz. zaprzeszły
(*c*, x_6): 3 os. lp., r. m., cz. przeszły + [*byłby*] = 3 os. lp., r. m., tryb przyp., czas przeszły
(*c*, x_7): 3 os. lp., r. m., cz. przeszły + [*byłbym*] = 1 os. lp., r. m., tryb przyp., czas przeszły
(*c*, x_8): 3 os. lp., r. m., cz. przeszły + [*byłbyś*] = 2 os. lp., r. m., tryb przyp., czas przeszły
(*c*, x_9): 3 os. lp., r. m., cz. przeszły + [*by*] = 3 os. lp., r. m., tryb przyp., czas nieprzeszły
(*c*, x_{10}): 3 os. lp., r. m., cz. przeszły + [*bym*] = 1 os. lp., r. m., tryb przyp., czas nieprzeszły
(*c*, x_{11}): 3 os. lp., r. m., cz. przeszły + [*byś*] = 2 os. lp., r. m., tryb przyp., czas nieprzeszły

(*d*, x_1): 1 os. lm., r. m., cz. przeszły + Ø (np. *pisaliśmy*)
(*d*, x_2): 1 os. lm., r. m., cz. przeszły + [*byli*] = 1 os. lm., r. m., cz. zaprzeszły

(*e*, x_1): 2 os. lm., r. m., cz. przeszły + Ø (np. *pisaliście*)
(*e*, x_2): 2 os. lm., r. m., cz. przeszły + F [*byli*] = 2 os. lm., r. m., cz. zaprzeszły

(*f*, x_1): 3 os. lm., r. m., cz. przeszły + Ø (np. *pisali*)
(*f*, x_2): 3 os. lm., r. m., cz. przeszły + [*będziemy*] = 3 os. lm., r. m., cz. przyszły
(*f*, x_3): 3 os. lm., r. m., cz. przeszły + [*będziecie*] = 3 os. lm., r. m., cz. przyszły
(*f*, x_4): 3 os. lm., r. m., cz. przeszły + [*będą*] = 3 os. lm., r. m., cz. przyszły

[15] Pominięte zostały tradycyjnie wyróżniane formy strony biernej i zwrotnej, bowiem przyjęto, że stanowią one konstrukcje składniowe, np. *jest wychowywany* – to związek formy czasownikowej z przymiotnikową; *myje się* – związek formy czasownika z formą biernikową zaimka zwrotnego.
[16] Pozostałe formy rodzajowe można wyprowadzić z form rodzaju męskiego za pomocą prostego algorytmu.

(*f*, x_5): 3 os. lm., r. m., cz. przeszły + [*byli*] = 3 os. lm., r. m., cz. zaprzeszły
(*f*, x_6): 3 os. lm., r. m., cz. przeszły + [*byliby*] = 3 os. lm., r. m., tryb przyp., czas przeszły
(*f*, x_7): 3 os. lm., r. m., cz. przeszły + [*bylibyśmy*] = 1 os. lm., r. m., tryb przyp., czas przeszły
(*f*, x_8): 3 os. lm., r. m., cz. przeszły + [*bylibyście*] = 2 os. lm., r. m., tryb przyp., czas przeszły
(*f*, x_9): 3 os. lm., r. m., cz. przeszły + [*by*] = 3 os. lm., r. m., tryb przyp., czas nieprzeszły
(*f*, x_{10}): 3 os. lm., r. m., cz. przeszły + [*byśmy*] = 1 os. lm., r. m., tryb przyp., czas nieprzeszły
(*f*, x_{11}): 3 os. lm., r. m., cz. przeszły + [*byście*] = 2 os. lm., r. m., tryb przyp., czas nieprzeszły

(*g*, x_1): bezokolicznik + Ø (np. *pisać*)
(*g*, x_2): bezokolicznik + [*będę*] = 1 os. lp., cz. przyszły
(*g*, x_3): bezokolicznik + [*będziesz*] = 2 os. lp., cz. przyszły
(*g*, x_4): bezokolicznik + [*będzie*] = 3 os. lp., cz. przyszły
(*g*, x_5): bezokolicznik + [*będziemy*] = 1 os. lm., cz. przyszły
(*g*, x_6): bezokolicznik + [*będziecie*] = 2 os. lm., cz. przyszły
(*g*, x_7): bezokolicznik + [*będą*] = 3 os. lm., cz. przyszły
(*g*, x_8): bezokolicznik + [*niech będzie*] = 3 os. lp., tryb rozk.

(*h*, x_1): 3 os. lp., cz. teraźniejszy + Ø (np. *pisze*)
(*h*, x_2): 3 os. lp., cz. teraźniejszy + [*niech / niechaj*] = 3 os. lp., tryb rozk.

(*i*, x_1): 3 os. lm., cz. teraźniejszy + Ø (np. *piszą*)
(*i*, x_2): 3 os. lm., cz. teraźniejszy + [*niech / niechaj*] = 3 os. lm., tryb rozk.

Wszystkie wyróżnione jednostki homonimiczne są wielofunkcyjne. Jeśli przy automatycznej segmentacji tekstu komputer je wyróżni (poda ich podstawową charakterystykę gramatyczną), muszą one zostać poddane dalszej procedurze rozpoznawczej, polegającej na poszukiwaniu w ich sąsiedztwie, w kontekście lewo- i prawostronnym, form leksemów BYĆ i NIECH, a w wypadku ich znalezienia – także na aktualizacji interpretacji morfologicznej[17].

Największą wielofunkcyjność wykazują formy wyrazowe leksemu BYĆ[18]. Mogą one pełnić następujące funkcje:
- samodzielnej formy wyrazowej, np. *jestem*;
- elementu formy wyrazowej składającej się ze słowa *jest* lub *był* i ruchomej cząstki osobowej, np. *jam jest*; *jam był*;
- elementu czasu przyszłego złożonego, np. *będę pisał*;

[17] Analityczne formy czasu przyszłego i przeszłego oraz trybu przypuszczającego tworzą także czasowniki niewłaściwe, np. *można było*, *trzeba będzie*, *warto byłoby*, nie będą one jednak tu analizowane.
[18] Widać to wyraźnie w artykule hasłowym BYĆ w SFWP.

– elementu czasu zaprzeszłego, np. *pisałem był.*

Słowo jako wykładnik dwóch form wyrazowych

Sytuacja, w której słowo stanowi wykładnik dwóch form wyrazowych, dotyczy:
– form wyrazowych przyłączających ruchome cząstki osobowe czasownika: *-m, -ś, -śmy, -ście*, np. *kiedyśmy (poszli), jam (to, jest), toś (zrobił)*, lub wykładnik trybu warunkowego *-by*, np. *żebyście (poszli)*[19];
– połączenia przyimka z morfemem *-ń*, np.: *zeń, weń*, a także powstałych w wyniku analogii *nań, zań, przezeń*;
– połączenia formy wyrazowej z partykułami *-ż(e)* lub *-li*, np. *dajże, kiedyż, znaszli.*

Morfemy osobowe aglutynacyjnych form czasu przeszłego i trybu warunkowego czasowników mogą się łączyć ortograficznie z inną jednostką tekstową, jak w zdaniach /6/-/8/. W wypowiedzeniu /9/ ten mechanizm jest obligatoryjny, por.:

/6/ *Wiem, żeście przeczytali książkę.*
/7/ *Wiem, żebyście przeczytali książkę.*
/8/ *Aleś powiedział!*
/9/ *Chcę, żebyście to zrobili.*

Końcówki osobowe mogą został przyłączone do większości części mowy (np. *rzecznikam, wreszcieśmy, dawnoście, zdrowaś*[20]). W wypadku zaimków nieokreślonych i pytajnych często zachodzi zjawisko homonimii[21]. Wykazuje się ono tu regularnością, która wynika z wielofunkcyjności morfemu *-ś*, będącego wykładnikiem 2 osoby lp. czasownika lub elementem zaimka nieokreślonego (*ktoś, jakiś* itd.). Na przykład słowo *coś* może zostać zinterpretowane jako – po pierwsze – zaimek nieokreślony (*coś*); po drugie – zaimek pytajny z ruchomą cząstką osobową (*co* + *ś*), por:

czegoś: (*a*, x_1): D. lp. COŚ
(*a*, x_2): zaimek pytajny z ruchomą cząstką osobową (*czego* + *ś*)
czemuś: (*a*, x_1): D. lp. COŚ
(*a*, x_2): zaimek pytajny z ruchomą cząstką osobową (*czemu* + *ś*)
czymś: (*a*, x_1): Msc. N. lp. COŚ
(*a*, x_2): zaimek pytajny z ruchomą cząstką osobową (*czym* + *ś*)

W taką homonimię wikłają się wszystkie formy wyrazowe należące do paradygmatu zaimków nieokreślonych zakończonych cząstką *-ś* odmiennych: *ktoś, jakiś, któryś* i nieodmiennych: *gdzieś, kiedyś, dokądś, ileś.*

Cząstka *-by*, którą uznajemy za morfem wchodzący w skład czasownikowych form wyrazowych, może ortograficznie zostać przyłączona do spójników (np. *alboby,*

[19] Dokładniej, słowo jest w tym wypadku wykładnikiem jednej formy wyrazowej i morfemu należącego do drugiej.
[20] Słowo *zdrowaś* można interpretować na dwa sposoby: jako połączenie formy przymiotnikowej *zdrowa* oraz albo końcówki osobowej *-ś* (*Zdrowaś już jest?*), albo formy 2 os. lp. czasu teraźniejszego czasownika BYĆ (*Zdrowaś, Maryjo, łaskiś pełna*), por. Z. Saloni, *Klasyfikacja gramatyczna leksemów polskich*, „Język Polski" LIV (1974), s. 3-13, 93-101.
[21] Por. U. Andrejewicz, 2001, *Polskie zaimki rzeczowne w ujęciu gramatycznym*, Białystok, s. 129-150.

chociażby, jeżeliby) i partykuł (np. *niechby, nużby, chybaby*). Takie złożone jednostki wyodrębnione podczas automatycznej analizy morfologicznej powinny zostać poddane dalszej analizie fleksyjnej.

Jeśli w procesie wyodrębniania wyrazów morfologicznych zostanie stwierdzona obecność czasownikowej formy 3 os. lp. lub lm. czasu przeszłego, w celu prawidłowego wyodrębnienia form wyrazowych należy szukać w najbliższym kontekście słów kończących się na *-m, -ś, -śmy, -ście* oraz *-by*. Wymienione wyżej homonimiczne formy czasu przeszłego czasownika należałoby zatem rozszerzyć o następujące:

(*c*, x_{12}): 3 os. lp., r. m., cz. przeszły + [*-(e)m*] = 3 os. lp., r. m., cz. przeszły
(*c*, x_{13}): 3 os. lp., r. m., cz. przeszły + [*-(e)ś*] = 3 os. lp., r. m., cz. przeszły
(*c*, x_{14}): 3 os. lp., r. m., cz. przeszły + [*-by*] = 3 os. lp., r. m., tryb przyp.

(*f*, x_{12}): 3 os. lm., r. m., cz. przeszły + [*-śmy*] = 3 os. lm., r. m., cz. przeszły
(*f*, x_{13}) : 3 os. lm., r. m., cz. przeszły + [*-ście*] = 3 os. lm., r. m., cz. przeszły
(*f*, x_{14}): 3 os. lm., r. m., cz. przeszły + [*-by*] = 3 os. lm., r. m., tryb przyp.

Analiza morfologiczna słów *zań, weń* itd. nie nastręcza trudności ani w segmentacji (można stworzyć po prostu listę takich jednostek), ani w charakterystyce morfologicznej, bowiem można tu jednoznacznie stwierdzić, iż słowa takie składają się z przyimka i pisanej zawsze łącznie formy *-ń*, należącej do paradygmatu zaimka ON.

Połączenia form wyrazowych z partykułami *-że, -li* muszą być oczywiście analizowane jako ciągi pisanych łącznie dwóch form wyrazowych, ale ich automatyczna analiza morfologiczna, ze względu na brak homonimiczności, jest dużo łatwiejsza[22]. Można oczywiście skonstruować zdania, w których omawiane jednostki będą uwikłane w homonimię, ale takie wypadki w istniejących korpusach są skrajnie rzadkie, por.: */10/ Kochali on ci mnie, czy nie kocha?* (z akcentem na sylabę ko).

Słowo jako składnik nieciągłej jednostki leksykalnej

Słowo może stanowić również składnik nieciągłej jednostki leksykalnej. Ze względu na to, iż postaci hasłowe jednostek leksykalnych są ustalane arbitralnie, ich listy mogą być różne w różnych ujęciach metodologicznych. Do takich jednostek na przykład zalicza się[23]:

- połączenia słów *co* i *jak* z przysłówkami i przymiotnikami w stopniu najwyższym, np. *co najmniej, jak największy, co najwyżej*;
- nieciągłe przysłówki, powstałe ze zleksykalizowanych wyrażeń przyimkowych (np. *z polska, po omacku, bez mała, na przekór, na schwał*) i inne ustabilizowane wyrażenia, np. *między innymi*[24];

[22] Jak zauważa R. Wołosz (2005, s. 29): „(...) o wiele łatwiejsze jest stworzenie algorytmu, dzięki któremu wyrazy takie [tj. np. *znaszli, chodźcież*] będą poprawnie analizowane, niż algorytmu wykrywającego analityczne (złożone) formy fleksyjne, np. *będą rysować, będą rysowały, byłbym zapomniał, spóźniło się*)".

[23] Por. SFWP, s. xxiii-xxiv.

[24] Wiele tego typu jednostek omawia M. Grochowski, 2002, *Wielowyrazowe jednostki funkcyjne. Wprowadzenie do problematyki*, [w:] *Problemy frazeologii europejskiej V*, Norbertinum, Lublin, s. 43-50. Jednostki tego typu w SFWP także zostały potraktowane jako leksemy nieciągłe. Sa

- czasowniki zwrotne, przy założeniu, iż strona nie należy do kategorii fleksyjnych (tzw. *reflexiva tantum*);
- formy imiesłowów przymiotnikowych (np. *znajdujący się*) i rzeczowników odsłownych (np. *pojawienie się*) z zaimkiem zwrotnym *się*;
- nazwy własne obcego pochodzenia zawierające pisaną osobno cząstkę typu *de, de la, du, la, le, van, von*, np. *de Gaulle, de la Roche, du Gard, la Fontaine, le Corbusier, van Gogh, von Jungingen*;
- stałe związki frazeologiczne, np. *zbić z pantałyku*.

Uznanie, że wymienione grupy zawierają leksemy analityczne (analityczne jednostki leksykalne), a nie konstrukcje składniowe, wymaga procedury, na podstawie której będą identyfikowane formy wyrazowe. Na przykład, ze względu na wysoką frekwencję w tekstach przyimków i zaimków wydaje się, że sprawdzanie ciągłości jednostek leksykalnych typu *po polsku* i *co najwyżej* powinno iść od członów *polsku* i *najwyżej*. To im należałoby przypisać dodatkowy poziom weryfikacji informacji morfologicznej. Z kolei stosunkowo rzadkie wystąpienia w tekście słów typu *von, de* itp. pozwala rozpocząć pogłębioną analizę morfologiczną właśnie od nich.

Segmentacja tekstu czasami jest związana z ortografią. Problem ten łatwo daje się przedstawić na przykładzie partykuły NIE[25]. Dotyczące jej pisowni reguły ortograficzne mają niebagatelny wpływ na segmentację. Partykuła NIE może być pisana na trzy sposoby:

a) rozdzielnie, np. nie czytam, nie najładniejszy
b) łącznie, np. niełady, niewesoło;
c) przez łącznik, np. nie-Afrykańczyk.

To, jak takie jednostki będą segmentowane, zależy od przyjętej konwencji. Ograniczmy na chwilę rozważania do wypadków (a)-(b).Tu możliwe są następujące rozwiązania:

- zarówno w ciągu *nie czytam*, jak i *niełady* występuje partykuła *nie* – wówczas słowo *niełady* jest wykładnikiem dwóch form wyrazowych;
- zarówno w ciągu *nie czytam*, jak i *niełady* słowo *nie* jest częścią formy wyrazowej – wówczas słowo *niełady* jest wykładnikiem jednej formy wyrazowej, a *nie czytam* należy uznać za formę analityczną;
- inaczej traktujemy *nie* w *nie czytam* (jako partykułę), inaczej w *niełady* (jako część formy wyrazowej).

Na marginesie warto zaznaczyć, że w nauczaniu szkolnym przyjmowane jest rozwiązanie trzecie, w wyniku którego formy tego samego przymiotnika czy przysłówka mają różne tematy fleksyjne, por. ***niełady***-*e* i *nie naj-**ładni**-ej*.

Z punktu widzenia automatycznej segmentacji tekstu wskazane problemy daje się łatwo rozwiązać. Ciąg *nie* łatwo można wydzielić: albo jest oddzielony z obu stron spacjami, albo z jednej strony spacją, z drugiej – łącznikiem, a przy pisowni łącznej również bez trudu wyodrębnimy go jako ciąg liter na początku słowa. Problem stanowi tu raczej

(1974) część z nich potraktował jako zwykłe związki składniowe, a te najbardziej zleksykalizowane, w których jedno ze słów występuje w tekście tylko w takim połączeniu, jako jednolite formy wyrazowe należące do leksemów nieodmiennych.

[25] Por. R. Wołosz, *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*, Warszawa 2005, s. 28.

decyzja, jak interpretować takie łącznie pisane *nie*. W słownikach pojawia się wiele haseł z *nie* pisanym łącznie. Notują one zatem takie regularnie tworzone jednostki, jak NIEAKTUALNY, NIEOBECNY czy NIEDUŻO itd., ale nie: NIE CZYTAĆ. To właśnie ma duży wpływ na wspomniane wyżej decyzje interpretacyjne[26].

Ciągi liter z łącznikiem również nastręczają wiele problemów segmentacyjnych. Po pierwsze, nie jest oczywiste, czy taką jednostkę traktować jako jedno słowo, czy dwa, ponieważ ten znak interpunkcyjny pojawia się w różnych typach jednostek, np. łączy człony niewystępujące oddzielnie i takie, które stanowią samodzielne jednostki, pojawia się ze względów graficznych itd., por. *cud-dieta* (ale: *dieta cud*), *laska-parasol*, *hokus-pokus*, *pseudo-Polak* (ale: *pseudoartysta*), *pół-Polka* (ale: *pół Polka pół Francuzka*), *XX-wieczny*.

Rygorystyczne opracowanie poruszonych w niniejszym teśkcie zagadnień jest niezbędne przy zautomatyzowanych pracach nad tekstem polskim. Artykuł stanowi propozycję usystematyzowania przydatnych w automatycznej analizie fleksyjnej informacji, które w wielu pracach językoznawczych pojawiały się, jednak nigdy nie były podstawowym przedmiotem opisu. Prezentacja problemów segmentacyjnych pojawiających się przy analizie morfologicznej pozwoliła na wykazanie zależności systemowych w występowaniu jednostek homonimicznych. Przeprowadzone rozważania pokazują także, że uzyskane podczas analizy morfologicznej wyniki zawsze są wypadkową przyjętych założeń metodologicznych dotyczących wyróżniania klas leksemów, ich paradygmatów oraz sposobu identyfikowania nieciągłych jednostek leksykalnych.

Wykaz skrótów

S	– klasa leksemów rzeczownikowych
A	– klasa leksemów przymiotnikowych
A(-adv)	– formy przymiotnikowe leksemów przymiotnikowych
A(adv)	– formy przysłówkowe leksemów przymiotnikowych
V	– klasa leksemów czasownikowych

Bibliografia

Andrejewicz U. 2001. *Polskie zaimki rzeczowne w ujęciu gramatycznym*. Wydawnictwo UwB, Białystok.

Awramiuk E. 2002. *Wpływ odstępstw od segmentacji ortograficznej na wyniki statystyczne "Słownika frekwencyjnego polszczyzny współczesnej"*. „Roczniki Humanistyczne" t. XLIX-L, 2001-2002 z. 6, s. 31-43.

[26] Oczywiście w automatycznej analizie tekstu pojawią się jakieś jednostkowe problemy. Na przykład w leksemach NIEBO, NIEMY, NIEBIESKI, NIEDOJDA itd. nie ma partykuły NIE; niektóre jednostki z *nie* mają znaczenie inne niż suma znaczeń partykuły i odpowiedniego leksemu, por. np. NIECZYSTOŚĆ, NIECHĘĆ; istnieją też takie leksemy, w których słowotwórczo daje się wyodrębnić przedrostek *nie-*, np. NIEZBĘDNIK (od *niezbędny*), NIEWOLNIK (od *niewolny*), ale czy można go sklasyfikować jako partykułę? Zupełnie na marginesie można wspomnieć o takich słowach, jak *najniebezpieczniej* czy *najniegrzeczniejszy*, gdzie *nie* (partykuła *nie*?) znajduje się w środku słowa. Przypadki te można jednak zdefiniować przez listę.

Bańko M. 2002. *Wykłady z polskiej składni.* Wydawnictwo Naukowe PWN, Warszawa 2002.

Bień J., Saloni Z. 1982. *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna).* „Prace Filologiczne" t. XXXI, s. 31-45.

Grochowski M. 2002. *Wielowyrazowe jednostki funkcyjne. Wprowadzenie do problematyki*, [w:] *Problemy frazeologii europejskiej V* pod red. A.M. Lewickiego. Norbertinum, Lublin, s. 43-50.

Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J. 1990. *Słownik frekwencyjny współczesnej polszczyzny*. PAN, Kraków.

Laskowski R. 1998. *Podstawowe pojęcia fleksji*, [w:] *Gramatyka współczesnego języka polskiego. Morfologia 1* pod red. R. Grzegorczykowej, R. Laskowskiego i H. Wróbla. Wydawnictwo Naukowe PWN, Warszawa, s. 125-150.

Saloni Z. 1996. *Homonimia a hasła w słownikach polskich.* „Język Polski" LXXVI 4-5, s. 303-314.

Saloni Z. 1974. *Klasyfikacja gramatyczna leksemów polskich.* „Język Polski" LIV, s. 3-13, 93-101.

Tokarski J. 1978. *Fleksja polska*. Wydawnictwo Naukowe PWN, Warszawa.

Wołosz R. 2005. *Efektywna metoda analizy i syntezy morfologicznej w języku polskim.* Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Wróbel H. 1995. *Problemy dyskusyjne w syntaktycznej klasyfikacji polskich leksemów*, [w:] *Studia gramatyczne XI*. Kraków, s. 7-18.

Streszczenie

Artykuł poświęcony jest omówieniu problemów segmentacyjnych, które pojawiają się podczas automatycznej analizy morfologicznej tekstu pisanego. Niepokrywanie się ortograficznej granicy słowa z granicą jednostki leksykalnej ma miejsce wtedy, gdy słowo stanowi jeden z elementów analitycznej formy wyrazowej, jest wykładnikiem dwóch form wyrazowych oraz kiedy okazuje się składnikiem nieciągłej jednostki leksykalnej. W artykule podano przykłady jednostek nieciągłych każdego typu, zasygnalizowano także zjawisko regularności pojawiania się w tekście jednostek homonimicznych.

Słowa kluczowe: segmentacja, jednostki nieciągłe, analiza morfologiczna