Scripta Neophilologica Posnaniensia, Tom VII, strony: 221– 239
Wydział Neofilologii, UAM Poznań, 2005

AUTOMATYCZNE METODY EKSCERPCJI NEOLOGIZMÓW, CZYLI SŁOWOTWÓRSTWO FAKTOGRAFICZNE

PIOTR WIERZCHOŃ

W przedstawianym w tym miejscu artykule zaprezentujemy konkretną metodę analizy językoznawczej, która służy do automatycznej ekscerpcji neologizmów[1]. Interesować nas zatem będą sposoby wykrywania nowych jednostek językowych. Na wstępie chcielibyśmy zaznaczyć, iż interesuje nas takie postępowanie, które będzie czytelne przede wszystkim w gronie językoznawców. To stwierdzenie może budzić w tym miejscu i w tej chwili zdumienie, jednak obserwując literaturę językoznawczą na przełomie XX i XXI wieku (np. czasopisma: *Computational Linguistics*, *International Journal of Corpus Linguistic, Literary and Linguistic Computing, Speech and Language Technology* etc.), trudno nie zauważyć, że czynności ekscerpcyjne, dokonywane np. dla celów leksykograficznych, w zasadzie wykonywane są przez inżynierów, w mniejszym stopniu przez językoznawców, filologów[2]. Tekst niniejszy będzie służył przedstawieniu metod wykorzystywanych zarówno w badaniach światowych, w mniejszym zakresie w badaniach krajowych, jak i metod, które opracowaliśmy, rozwinęliśmy i testowaliśmy we własnym zakresie.

Skoncentrujemy się na następujących trzech pytaniach, które pomogą uszczegółowić problematykę pracy:

---

1 W sprawie teminu *neologizm* por.: Buttler 1962, Kurkowska 1956, Smółkowa 2001, Stoberski 1976, Waszakowa 1997, Wawrzyńczyk 1992, Wawrzyńczyk 1993, Zagrodnikowa 1982.
2 Por.: Jadacka 2001, Smółkowa 2001, Waszakowa 2001, Wawrzyńczyk 2000.

a) jakich obiektów poszukujemy,
b) gdzie poszukujemy tych obiektów oraz
c) w jaki sposób ich poszukujemy?

W odpowiedzi na pytanie pierwsze należy powiedzieć, iż o ile w latach 70., 80. czy nawet jeszcze w latach 90. prace ekscerpcyjne koncentrowały się na próbach ekscerpcji jednostek ciągłych, a więc elementów językowych traktowanych od spacji do spacji (ekscerpcja głównie słów), lata ostatnie przynoszą już bardzo zaawansowane koncepcje teoretyczne i ich praktyczne weryfikacje, w których omawia się możliwości ekscerpcji jednostek bardziej skomplikowanych syntaktycznie (kolokacje)[3]. Skrajnym przejawem analiz w tym nurcie badawczym jest określenie formalno-technicznych kryteriów ekscerpcji idiomów, które, jak wiadomo, najtrudniej poddają się jakiemukolwiek formalnemu, rygorystycznemu opisowi[4].

Gdy określono już kryteria segmentacji obiektów, które chcemy poszukiwać, w naszym wypadku będą to słowa-neologizmy, kolejnym krokiem będzie znalezienie odpowiedzi na pytanie: gdzie możemy poszukiwać interesujących nas obiektów? To być może nieistotne z punktu widzenia samego procesu ekscerpcji pytanie uznajemy za fundamentalne z tego powodu, że w materiałowej pracy lingwistycznej językoznawca i tak na pewnym etapie procesu ekscerpcji będzie musiał rozstrzygnąć problem, skąd pozyskiwać materiał dla analiz lingwistycznych: czy np. a) czynić to poprzez skanowanie książek, b) pozyskiwać dane z Internetu[5], c) przenosić mikroklisze na twardy dysk w formacie PDF bądź TXT, XML itp., czy wreszcie: pozyskiwać teksty w jeszcze inny, powyżej nie wymieniony sposób[6].

Jeżeli na uwadze mieć korzyści płynące z zastosowania automatycznych metod ekscerpcji, to należy tu powiedzieć przede wszystkim o przyroście nowego materiału dla zastosowań lingwistycznych oraz o uporządkowaniu wiedzy o języku (w interesującym nas fragmencie). W pierwszym wypadku należy powiedzieć, iż to właśnie nie w roku 1994[7], czyli zaledwie 10 lat temu, nie w roku 1995, ale raczej od roku 1997, w związku z rozwojem procesorów klasy Pentium, następuje gwałtowny przyrost materiału tekstowego możliwego do wykorzystania w badaniach językoznawczych. Ale dopiero przełom milenijny przynosi niebywałą ekspansję tekstów w postaci elektronicznej (e-książki, archiwa internetowe itp.). Bez przesady można stwierdzić, iż obecnie ludzkość dysponuje takimi tekstowymi danymi lingwistycznymi, jakimi nie dysponowała od początku posługiwania się pismem. Z drugiej strony należy stwierdzić, iż niemożliwe jest automatyczne operowanie danymi językowymi bez możliwie rygorystycznego sprecyzowania ich własności. Inaczej mówiąc: aby efektywnie poszukiwać konkretnych jednostek językowych, należy możliwie dokładnie określić ich inwarianty graficzne, a więc te elementy, które będą stałe, niezmienne. Ta techniczna potrzeba z kolei wymusza sformułowanie konkretnych warunków lingwistycznych, jakie musi spełniać poszukiwana jednostka[8].

W dalszej kolejności przestawimy przygotowania konieczne do przeprowadzenia operacji automatycznego wyszukiwania neologizmów. Po pierwsze, musimy dysponować określonym zbiorem tekstów; liczba tych tekstów nie jest istotna, tj. może to być dowolna liczba tekstów. Ich wielkość także w prezentowanej metodzie nie ma znaczenia, może to być np. powieść „Anna Karenina", wiersz Mirona Białoszewskiego, a także notatka za spotkania sportowego lub notatka z dowolnego spotkania. Kolejno ze zgromadzonych tekstów wypisujemy wszystkie słowa (wyliczając je kolejno na schemacie: słowo 1 – słowo 10). To listowanie odbywać się będzie w sposób automatyczny. Tę pracę wykonuje komputer.

3 Por. natomiast jedną z pierwszych prób automatycznej ekscerpcji kolokacji w: Jones, Sinclair 1974.

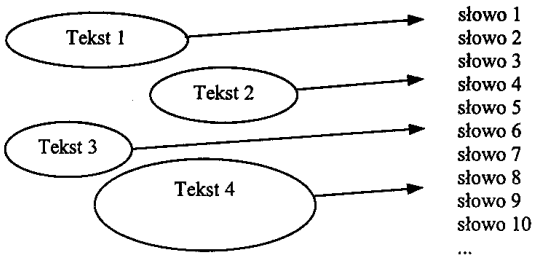
4 „Our aim is to provide a procedure that would enable us to **retrieve automatically** [podkr. – P.W.] idiomatic expressions from large text corpora" (Degand, Bestgen 2003: 250); por. Čermák 2001, Stubbs 2002.

5 Por. metodę: Sgarbas et al. 2003; dla języka polskiego por. pracę A. Buczyńskiego pt. *Pozyskiwanie z Internetu tekstów do badań lingwistycznych* (Praca wykonana pod kierunkiem dra hab. Janusza S. Bienia; Katedra Lingwistyki Formalnej UW).

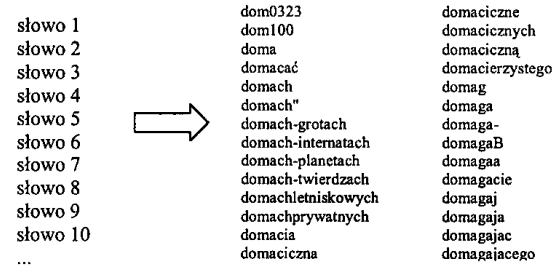
6 Zazwyczaj: powodując konsekwencje prawne, związane np. z ochroną praw autorskich; por. Pawłowski 2003.

7 Por. dane ilościowe w: Stubbs 1996: 231 – 234, Svartvik 1992; zwłaszcza: Gilquin 2002: 191 – 195.

8 Por. prace (ćwiczenia) z danymi korpusowymi (tj. formułowanie inwariantów graficznych poszukiwanych jednostek leksykalnych) zaprezentowane w pracach Bańki (2001, 2002); por.: Daciuk 1999, Graliński 1998, Graliński, Krynicki 2000, Koval et al. 2000, Kučera 2002, Obrębski 1998, Oflazer 1996, Sgarbas et al. 2003, Suszczańska et al. 2000, Wypych 2002, Vetulani et al. 1998.



Schematyczna lista słów (na schemacie: słowo 1 – słowo 10) w rzeczywistości może mieć postać następującą: *dom0323, dom100, doma, doma-, domacać, domach, domach", domach-grotach* itp.



Przedstawiana w tym miejscu lista to lista pobrana z własnych zasobów tekstowych. Widzimy na tej liście zarówno jednostki poprawne, akceptowalne, np. *domacać, domaciczna, domagać, domagająca*, jak i jednostki, które są po prostu błędami: typograficznymi lub ortograficznymi (np. *dom0320, domaga-, domagaa, domagajaca* itd.).

W kolejnym kroku przedstawiamy ideę analizatora morfologicznego. Analizator morfologiczny to program komputerowy pozwalający danemu słowu przyporządkować jego własności morfologiczne (gramatyczne)[9]. Analiza morfologiczna wykonana automatycznie umożliwia dostęp do danych w formacie niezwykle przydatnym dla dalszych badań językoznawczych. Przede wszystkim analiza pozwala zaobserwować słowa wraz z ich możliwymi charakterystykami gramatycznymi, czasem także wraz z ich analizą morfemową (podział na temat i końcówkę). Już w tym punkcie można zadać pytanie: czym miałaby być owa charakterystyka? Charakterystykę taką stanowi zbiór kategorii gramatycznych właściwych danej części mowy[10]. Zatem jeżeli przyjmiemy, że np. dla rzeczownika jesteśmy w stanie określić wartość kategorii: Przypadek, Rodzaj oraz Liczba, to charakterystyką analizowanego słowa *długopisem* będą cechy – Narzędnik, Rodzaj męski oraz Pojedynczość.

9 W chwili opracowywania tekstu automatyczna analiza morfologiczna jest praktyką stosunkowo powszechną, dla której można odnaleźć bogatą literaturę przedmiotu (por. cenny ze względu na dane historyczne artykuł: Bień, Szafran 2001). U zarania prac automatyzujących polską analizę morfologiczną najistotniejszą rolę odegrali Z. Saloni (m.in. poprzez opracowanie SIAT, a także poprzez promocję niezwykle istotnych w krajowym językoznawstwie prac: Szafran 1993, Wołosz 2000) oraz J.S. Bień. Por. także np.: Kaalep 1997, Koval et al. 2000, Proszeky, Tihanyi 1992, Suszczańska et al. 2000, Vetulani et al. 1998.

Przechodząc do konkretnych analiz morfologicznych, pierwszej przykładowej analizie będziemy poddawać pięć słów: a) *dom*, b) *domek*, c) *antydom*, d) *eurodom*, e) *euurodom**. Ostatnie słowo jest słowem celowo zapisanym z błędem, aby ukazać, w jaki sposób będzie przebiegać analiza morfologiczna dla jednostki wadliwej. Zatem dokonujemy analizy słów *dom* i *domek*. Analiza[11] ta wskazuje, że:

Analysis of "dom":
dom[Sm3]+[SG]+[NOM]
dom[Sm3]+[SG]+[ACC]

10 „Przez *analizę morfologiczną* rozumiemy pewną operację (program lub algorytm ją realizujący nazwiemy *analizatorem morfologicznym*), która dla każdego słowa stanowiącego dane wejściowe produkuje pewien jego opis" (Bień, Szafran 2001: 171).

11 Prezentujemy format opracowany przez R. Wołosza (Wołosz 2000). Format opracowany przez K. Szafrana (Szafran 1993) prezentuje się następująco:

dom {{(N) < dom(mIV::m3)+ } }%
domek {{(N) < domek(mIII::m3)+ } }%
nierozpoznane: antydom { }, eurodom { }, euurodom { }.
Format opracowany przez TiP Sp. z o.o. prezentuje się następująco:

dom dom[Sm3]+[11,41]
domek domek[Sm3]+[11,41]
nierozpoznane: antydom, eurodom, euurodom.

Analysis of "domek":
domek[Sm3]=dom+ek[SG]+[NOM]
domek[Sm3]=dom+ek[SG]+[ACC]

dla słowa *dom* widzimy dwojaką interpretację: rzeczownik męski [Sm3], w liczbie pojedynczej [SG] oraz w przypadku mianownika [NOM] lub biernika [ACC]. Podobnie dla słowa *domek*, przy czym w tym wypadku uwidacznia się podział morfologiczny na temat *dom-* i końcówkę *-ek*. Można także założyć, że konkretny analizator dysponuje bogatszymi regułami analizy, tj. pozwala na poprawną analizę słów hybrydalnych, tj. np. z określonymi prefiksami, np.:

Analysis of "antydom":
anty[sub']+dom[Sm3]+[SG]+[NOM]
anty[sub']+dom[Sm3]+[SG]+[ACC]

Analysis of "eurodom":
euro[sub']+dom[Sm3]+[SG]+[NOM]
euro[sub']+dom[Sm3]+[SG]+[ACC]

Analiza ta ukazuje, iż możliwe jest rozpoznanie istniejącej w zbiorze tekstów formy graficznej *antydom* i *eurodom* jako form: rzeczownik męski [Sm3] w liczbie pojedynczej [SG] oraz w przypadku mianownika [NOM] lub biernika [ACC]. Dodatkowo uwidacznia się analiza prefiksacji w postaci: *anty-*, *euro-* (symbol: [sub'] jest symbolem kodu analizatora i oznacza prefiks w analizowanej formie).

Na tak naszkicowanym tle należy zapytać, jaki wynik analizy morfologicznej uzyskamy dla słowa *euurodom*? Czy analizator potrafi zanalizować taką formę? Otóż w kolejnym kroku przedstawione zostaną przypadki nierozpoznania słowa. Inaczej mówiąc: interesuje nas to, jak zachowa się analizator morfologiczny, gdy nie będzie potrafił rozpoznać danego słowa.

Jeżeli analizator nie rozpozna słowa, to w tym wypadku mamy do czynienia z dwojaką interpretacją:

a) analizowane słowo może zawierać błąd ortograficzny, typograficzny itp.,
b) analizowane słowo może być neologizmem, a więc nowym słowem wprowadzonym do języka, do tekstu.

Zatem w celu efektywnej ekscerpcji neologizmów interesuje nas tylko stan, w którym dane słowo nie zostało przez analizator zinterpretowane morfologicznie; mówiąc innymi słowy: analizator nie zna danej jednostki języka, nie potrafi jednoznacznie przyporządkować własności gramatycznych neologizmowi. Stąd jednostki, przykładowo[12]: *pseudorozwiązanie, kontrdecyzja, świtoniada, prokorupcyjny, debiliada, superpasywny, pseudokombatanctwo, superdotacja* można interpretować jako potencjalne neologizmy. W dalszej kolejności przedstawimy analizy morfologiczne dokonane w celu ekscerpcji neologizmów dla języków: rosyjskiego, koreańskiego oraz angielskiego. Celem tych analiz będzie poszerzenie istniejących list słownictwa, np. w celu wykorzystania w leksykografii. Zatem interesuje nas taka analiza morfologiczna, która prowadzi do ekscerpcji nienotowanych wcześniej jednostek języka.

Dla języka rosyjskiego plikiem wejściowym będzie plik na który składają się przykładowo formy: *счастливый, суперсчастливый, суперсчастлива, суперсчастливых*. Z podanych form analizator rozpoznał jedynie *счастливый*. Nierozpoznane[13] pozostały trzy formy: *суперсчастливый, суперсчастлива,*

12 Kolejne przykłady (wybór z kilkudziesięciu tysięcy): *aerogryf, antyantyantyantyantyantyantyrakieta, antybolszewik, antyhacker, antykryminał, antymanipulacyjny, antymarksowski, antyogniowy, bezklejowy, bezkonkursowy, bezprzesadny, bezszczękowy, całościenny, czarnoszafirowy, ćwierćromans, drobnoguzkowy, ekofilm, ekopartner, eksgenerał, eksgubernator. elektrobudowa, elektrodyfuzor, eurodata, euroeuforia, fluoropochodny, fluorowodorek, fototerapeutyczny, hipertoń, izopentan, kontrgatunek, kryptoideologiczny, makrodramat, megakompleks, megaobrót, megapaństwo, metaamfetamina, metahipostaza, metapustka, międzybór, międzyświatowy, mikromanipulacja, mikromiasteczko, ministówa, minitraktat, monocło, niskowęglowodanowy, nowoworoneski, okołostrefowy, paleoteropod, paraospowy, pięciopalec, podchlor, poddozorca, podlegnicki, ponadobyczajowy, postsyjonizm, pozajęzyczny, półatomowy, półmodernizm, półosłona, półwiorstowy, pradziadunio, prapokolenie, proalgierski, promazowiecki, przeciwcarski, przeciwwypływowy, pseudoburżuj, pseudosondaż, samoodcięty, samoorganizacyjny, samouzdrowienie, schizoencefalopatia, srebrnoświetlny, starogrodzki, stereoselektywny, stutytułowy, subedycja, subkrytyczny, superciało, superdyplomatyczny, superistotny, supersłoneczny, superśmigłowiec, superterminal, szaropopielaty, technomagia, telematematyk, telerandka, telezabawa, wewnątrzregionalny, wielotypowy, wszechwładnący, wszechżyto, zachodnioaramejski.*

13 Aby unaocznić sens ekscerpcji neologizmów rosyjskich poprzez analizę morfologiczną, przedstawiamy dla przykładu niewielką (wybór z kilkunastu tysięcy otrzymanych w ten sposób neologizmów) serię wybranych form nierozpoznanych przez analizator: *абсорбированность, авиагорноспасатель, авиамотомехборьба, автобиографичеокий, автобытавтоматик, автогностический, автодетектировать, автокомпрессорный, автокорректировочный, автомагнетический, автомобилизировать, автопринадлежность, автопроблематика, автопроецирование, авторегенерационный, автороцентрический, автостерилизатор, автоэволюционизм, автоэволюционист, агитаторствовать, адмиральствовать, айсбергоподобнымн, аквариумообразный, аккредитованность, аккуратноуложенный, амбразуродозорный, аметистоподобной, аметистоподобный, ангелоподнобность, англинизированность, англофильствующий, англошотландский, антиабстракционист, антиакселерационистка, антиантропогенный, антиаристократичный, антиархаистический,*

суперсчастливых. Właśnie te trzy formy należy poddać w dalszej kolejności analizie manualnej, prowadzącej w efekcie do ekscerpcji neologizmów. Analizator bowiem określa te trzy słowa jako nieznane (informacja -1 0; każda inna informacja, np. 158606, jest oznaczeniem kodu gramatycznego w danym analizatorze). Powstaje w tym miejscu pytanie: co lub kto decyduje o tym, że dane formy są rozpoznawane lub nie przez analizator. Co stanowi podstawę do określania neologizmów, co jest monitorem dla nowego słownictwa? W przypadku języka rosyjskiego może to być *Грамматический словарь русского языка* (Андрей Анатольевич Зализняк). Oznacza to, że jeżeli analizator będzie analizował tekstową (np. fleksyjną) formę, która została ujęta w tym słowniku w postaci kanonicznej, wówczas tę formę zaakceptuje, a jeżeli analizie będzie poddawana forma wcześniej niezanotowana w przykładowym gramatycznym słowniku, to analizator zwróci odpowiedź: „forma nierozpoznana". Dla analizatorów języka polskiego podstawę materiałową stanowi najczęściej *Słownik języka polskiego* pod redakcją W. Doroszewskiego. Naturalnie, zbiór haseł z tego słownika ujęty w danym analizatorze bardzo często uzupełniany jest materiałem z nowszego

антибизнесовский, антибиостимулятор, антиблевательный, антивегетарианец, антививисекционистский, антиврангельский, антивысокомерный, антигероический, антигосударственный, антигосударственныя, антигоударственный, антигригорьевский, антиджасперианцевый, антидиккенсовский, антидомициановский, антидэвионовский, антиескопальный, антизаинтересованность, антизоммервильдовец, антиинфраустройство, антиистеблишментный, антикапонистский, антикасталийский, антиклаустофобия, антиконтрреформационный, антикриптографический, антикюстиновский, антилеконтовский, антиличностность, антиматематичность, антиматериальность, антиматоборовский, антиматримониальный, антиморганистский, антишредингеровский, антономатический, антроисследование, антрополингвистик, антропосоциальный, антропосоциогенез, антропотеический, антропофагический, антропофанический, аполлонообразный, апостолоподобный, аргентинизировать, аристократничать, артспецшкольниковый, архангелоподобный, археоисследователь, археологизировать, археологоразведка, архибессовестный, архиблагоглупость, архибюрократический, архивоинственный, архивпечатляющий, архидемонический, архидубльнаигениальнейший, архикардинальный, архикатолический, архинепримиримый, архиоппортунистический, архипредставительный, архипренеприятный, архипренеприятный, архипреступность, архипрозаический, архипроникновенный, архиразочарование, архиреволюционный, архисимволический, архифедеративный, архичрезвычайный, асинхронизировать, астрогеотектоника, астроизотерический, астрокоординатор, астронавигапионный, астронавигаторство, астронавигаторша, астрополитический, астроцентрический, асфальтированнный, асфальтодробилка, асфальтопихатель, атоллокоралловый, атомомолекулярный, атомомолекулярный, аудивидеоаппарат, аудиовидеомонитор, аудиовидеопроизведения, аудиоимагинативность, аудиокоммуникация, аудиокоммуникационный, аудиосканирование, аудиоэкранировка, ауропроекционный, афромусульманский, аэроакробатический, аэрокартографический, аэрокондиционирование, аэрофотолаборатория.

Słownika języka polskiego pod redakcją M. Szymczaka, a sporadycznie także materiałem z najnowszych wydawnictw leksykograficznych, poprawnościowych, np. ze *Słownika ortograficznego* PWN[14].

14 Braki w najdoskonalszych nawet słownikach powodują, iż autorzy analizatorów często samodzielnie umieszczają w bazach analizatorów jednostki bezpośrednio ekscerpowane z tekstów. Pod tym względem imponować może bogactwo leksykalne (zwłaszcza ze względu na aktualność słownictwa) analizatora TiP. Sp. z o.o. Naturalnie, możliwa jest ewentualnie (o ile jednostek tych nie ma w bazie podstawowej analizatora) implementacja algorytmu pozwalającego na generowanie form przysłówkowych z przymiotników: *bebechowato, ekspresyjnie, embriologicznie, insurekcyjnie, mitotwórczo, psychologistycznie, psychometrycznie, semiotycznie, sokratycznie* czy form gerundialnych typu: *sprawozdawania, bumblowanie* itd. Nadal jednak pozostaje w oczach użytkownika analizatora uznanie dla ujętych w algorytmie postaci: *michalikowy, nadświadomy, niesprzedawalny, odmóżdżony, podłatusiale*. Tu dodatkowo należy podkreślić dwie kwestie. Pierwsza, to ewentualna możliwość konkatenowania elementów bazy słownikowej analizatora: *antyelektrostatyczny, antyimperialistyczny, antyintelektualny, antykapitalistyczny, antykoincydencyjny, antymonarchistyczny, antynaturalistyczny, antypozytywistyczny, antyreformatorski, antyrepublikański, antyrojalistyczny, antysocjalistyczny, antysyjonistyczny., antyterrorystyczny, autoradiograficzny, beztransformatorowy, bioelektromagnetyczny, elektrohydrauliczny, elektroinstalacyjny, elektroluminescencyjny, elektrometalurgiczny, elektropneumatyczny, elektroprzewodnictwo, elektrostymulacyjny, elektroterapeutyczny, intersubiektywizacja, jednokondygnacyjny, jednowodorotlenowy, kontrrewolucjonistka, magnetohydrodynamiczny, międzybiblioteczny, międzycywilizacyjny, międzycząsteczkowy, międzykrystaliczny, międzyspółdzielczy, międzyśrodowiskowy, międzyzrzeszeniowy, neoimpresjonistyczny, niskoczęstotliwościowy, niskokwalifikowany, niskotemperaturowy, ogólnometodologiczny, pięciocentymetrowy, polichlorowinylowy, południowoamerykański, południowoazjatycki, południowoindyjski, południowokoreański, południowoniemiecki, ponadnarodowościowy, postkapitalistyczny, postmodernistyczny, postsocjalistyczny, północnoamerykański, północnowietnamski, przeciwastmatyczny, przeciwwybuchowy, przeciwuderzeniowy, przedmarksistowski, pseudofilozoficzny, pseudointelektualny, pseudosocjologiczny, pseudostereofoniczny, psychoneurologiczny, psychosocjologiczny, socjolingwistyczny, spektrofotometryczny, szerokoprzestrzenny, średniotemperaturowy, środkowoamerykański, środkowoeuropejski, teleinformatyzacja, termoluminescencyjny, trójprzymiotnikowy, ultrakonserwatywny, wczesnorenesansowy, wewnątrzśrodowiskowy, wielkolaboratoryjny, wielodyscyplinarny, wielopierścieniowy, wielospecjalistyczny, wielowodorotlenowy, wschodnioafrykański, wschodnioamerykański, wschodnioatlantycki, wschodnioazjatycki, wschodnioberliński, wschodniogermański, wschodnioniemiecki, wschodniosłowiański, wysokoczęstotliwościowy, wysokokwalifikowany, zachodnioafrykański, zachodnioamerykański, zachodnioaustralijski, zachodnioazjatycki, zachodnioberliński, zachodniobiałoruski, zachodniosłowiański, zachodniosyberyjski*. Odmiennym zagadnieniem jest bardzo pomocna możliwość automatycznej, rekurencyjnej konkatenacji liczebnikowych, typu np.: *czterdziestodziewięcioprocentowy, czterdziestopięciomiliardowy, czterdziestoześciotysięczny, czterystupięćdziesięciostronicowy, czterystusześćdziesięcioosobowy, dwudziestopięciocentymetrowy, dwudziestosiedmiokilometrowy, dwustukilkudziesięcioosobowy, dwustupięćdziesięciokilogramowy, dwustupięćdziesięciotysięczny, dwustusześćdziesięciotysięczny, dziewięćdziesięciocentymetrowy, dziewięćdziesięciodwustronicowy, dziewięćdziesięciokilkuprocentowy, dziewięćdziesięciomiliardowy, dziewięćdziesięcioośmioletni, dziewięćdziesięciopięcioletni, dziewięćdziesięciopięcioosobowy, dziewięćdziesięciopięcio-*

W dalszej kolejności zaprezentowana zostanie analiza materiału koreańskiego[15]. Analizie[16] poddano osiem form: 가격, 가넷, 가는, 가능, 가능율, 가능점, 가능한, 가능항. W dwu przypadkach (2 oraz 8) analizator wskazał na błędność formy, tj. jako wynik wskazał odpowiedź: 수정필요, co oznacza: „do poprawienia". Manualna analiza formy 2 oraz 8 dostarcza jednak nowych faktów: otóż forma 2 jest formą faktycznie interesującą z językoznawczego punktu widzenia (jest to stosunkowo rzadki termin techniczny), jednak forma 8 jest formą błędną, jest to przykład błędu literowego. Sukcesem zatem w tych analizach jest tylko[17] automatyczna ekscerpcja formy 2.

---

procentowy, dziewięćdziesięciopięciotysięczny, dziewięćdziesięciotrzymetrowy, osiemdziesięciodziewięcioletni, osiemdziesięciopięcioosobowy, pięćdziesięciodziewięcioletni, pięćdziesięciosiedmiometrowy, pięćdziesięcioszesćioipółcalowy, pięćdziesięcioszesćioprocentowy, pięćdziesięcioszesćiotysięczny, siedemdziesięciocentymetrowy, siedemdziesięciodziewięcioletni, siedemdziesięciojednostronicowy, siedemdziesięciopięciocentymetrowy, siedemdziesięciopięciogramowy, siedemdziesięciopięcioosobowy, siedemdziesięciosiedmioletni, siedemdziesięciosiedmiostronicowy, siedemdziesięcioszesćioletni, siedemdziesięcioszesćioprocentowy, siedemdziesięciotrzytysięczny, studwudziestoszesćiostronicowy.

15 W kwestii analizy języka koreańskiego por. ujęcie Kang, Kim 1994, Lee et al. 2003, Park et al. 1997, 1998. W kwestii etykiet gramatycznych (NNG, XSN, XR itd.) por. literaturę w poz. Han et al. 2002. Badania morfologiczne oraz analizę n-gramową por. w: Kwon, Park 2003.

16 Wynik analizy:

가격 가격/NNG 규칙
가넷 가넷/NF 수정필요
가는 가/VX+ 는/ETM NULL
가능 가능/NNG 규칙
가능율 가능/XR+ 율/XSN NULL
가능점 가능/XR+ 점/NNG NULL
가능한 가능/NNG+ 하/XSA+ ㄴ/ETM 규칙
가능항 가능항/NF 수정필요

17 Por. przykładowo inne formy zakwalifikowane jako nieznane analizatorowi: 게이트웨이, 그룹와이즈, 극초대규모, 기가헤르츠, 나노프로그래밍, 나이퀴스트, 내로캐스트, 내비게이션, 내비게이터, 내비게이트, 네오코그니트론, 네츠페디션, 네트비에스디, 네트커머스, 다이나로드, 다이렉트드로, 다이렉트비디오, 다이렉트엑스, 다이렉트인풋, 다익스트라의, 데이터그램, 덴드라이트, 디멀티플렉서, 디스어셈블링, 디스크램블러, 디스테이징, 디스파이킹, 디지사이퍼, 디지타이징, 디프래그먼테이션, 디플러션형, 라이노트로닉, 라이트사이징, 랩어라운드, 로크레머틱스, 로토스코프, 리플렉티브, 릴라이어블, 마이크로머시닝, 마이크로벤드, 마이크로세그먼테이션, 마이크로스프트, 마이크로초, 마이크로컴, 마이크로퍼블리싱, 마이크로폼, 마이크로필르머, 마이크로필르밍, 매시매티카, 맥바이너리, 맥클루스키법, 맨들브로트, 멀티바이브레이터, 멀티플렉서, 멀티플렉싱, 메리디오널, 메타크롤러, 모더레이터,

Dla języka angielskiego poddano analizie następujące przykłady: *urbanwear, echolocate, dollarization, strawberry, strawbery**. Celowo w ostatnim przykładzie, tj. dla jednostki *strawbery* wprowadzono błąd literowy, by przetestować zachowanie analizatora. Wynikiem analizy były dwie formy rozpoznane: *echolocate, strawberry*, zatem te formy nie będą nas w dalszym ciągu interesowały, natomiast trzy pozostałe słowa: *urbanwear, dollarization* oraz *strawbery* w ramach przedstawianej tu metody musimy poddać ponownej – manualnej – analizie. Z kolei wynikiem tej analizy jest następujący wniosek: z trzech form wejściowych *urbanwear, dollarization* oraz *strawbery* tylko dwie pierwsze są neo-

---

모듈레이티드, 몬스터보드, 바이오메트릭, 바이오센서, 바이오컴퓨터, 백프로퍼게이션, 베이퍼웨어, 보더매니저, 브레드보드, 브로셔웨어, 브리게이드, 브리지웨어, 브이디아이, 비디오그램, 사이리스터, 사이버벨럭스, 사이버서퍼, 사이브레리언, 서브스레솔드, 서브엘리먼트, 세그먼테이션, 슈퍼스케일러, 스칼라빌리티, 스캐니메이션, 스캐니메이트, 스컬프처형, 스크래치패드, 스크램블러, 스트림웍스, 스플라이싱, 스플리터가, 시그모이드, 시뮬레이티드, 시시더블유, 시프트클럭, 아로마웨어, 아이네타즌, 아이솔레이터, 아이엔에스, 안티스패머, 안티에일리어싱, 알타비스타, 알파지오메트릭, 애스터리스크, 애크로배트, 애트리뷰트, 액티베이션, 앤스로포모픽, 앰플리튜드, 어그리먼트, 어렌지먼트, 어세스먼트, 어소시에이트, 어시밀레이터, 언블라인드, 언인스톨러, 언카탈로그, 얼라인먼트, 에듀파인더, 에라토스테네스의, 에스더블유, 에이브이알, 에일리어네이션, 에일리어싱, 에코플렉스, 에피택셜법, 엔에프에스, 엔터프라이즈, 엘라스토머릭, 오거나이제이션, 오리엔티드, 오브젝티브, 옴니포인트, 웨어하우징, 웹크라울러, 유니프로세서, 이미디어트, 이오나이저, 이퀄라이저, 이프덴엘스, 익스체인지, 익스플로더, 인스트루먼츠, 인터레스트, 인터리피터, 인텔리전트, 인텔리전트패드, 인트라넷웨어, 인트라빌더, 인티그레이션, 인포미디어리, 인핸스먼트형, 자이노이드, 제너레이터, 제로그래피, 지오메트릭, 카디널리티, 카운터프로퍼게이션, 캐스케이딩, 캘리그래픽, 커넥셔니스트, 커넥셔니즘, 커뮤니케이션즈, 컨피그시스, 컴포넌트웨어, 컴포넌트형, 컴퓨니케이션, 컴플라이언스, 코그니트론, 코포레이트, 코프로세서, 코히어런스, 코히어런트, 크라이오트론, 크로마키잉, 크리플웨어, 클러스터링, 클리어링하우스, 클리어비전, 클럭드로윌, 클린스위프, 타이포그래피, 타이프메틱, 테라바이트, 테라헤르츠, 테이블웨어, 테크노네트워크, 테크노마, 테크노파크, 테크노폴리스, 텍스트에디트, 텍트로닉스, 텔러매틱스, 텔레가이드, 텔레데이터, 텔레라이팅, 텔레오토그래프, 텔레인포매틱스, 텔레커뮤니케이션즈, 텔레포스트, 텔로토그래프, 트라이어드, 트랜스폰더, 트랜스퓨터, 파플레이트, 파플레이티드, 패스트이더넷, 퍼셉트론즈, 퍼스펙티브, 페이크웨어, 페타바이트, 포스트앱블, 포토크로믹, 포토텔레그래피, 프라운호퍼, 프라이스캡, 프래그매틱스, 프로토타이핑, 프로파일러, 프로파일링, 프리미티브, 프리어셈블, 플로차팅법, 하이버네이션, 하이퍼링크, 하이퍼큐브, 하이퍼터미널, 하이퍼토크, 하이퍼패드, 핸드셰이킹, 헤테로다인, 홀로그래프, 홀로그래피; przy czym użycia: 내로캐스트, 내비게이션, 내비게이터, 내비게이트, 마이크로벤드, 마이크로스프트, 멀티플렉서, 익스플로더, 파플레이트, 파플레이티드 można uznać za wadliwe ortograficznie.

logizmami, natomiast forma *strawbery* to zgodnie z oczekiwaniem przykład błędu literowego – usterki graficznej. Podsumowując: analizę automatyczną w tym przypadku cechuje stosunek (sukces) 2 do 1 (dwie formy wyekscerpowane jako neologizm, natomiast jedna forma wyekscerpowana jako błędna, wadliwa)[18].

Przechodząc do podsumowania metody, chcielibyśmy zwrócić uwagę na jej zalety i wady. Przede wszystkim należy powiedzieć w tym wypadku o zalecie, jaką jest szybki, natychmiastowy dostęp do wyniku. Otóż analizy tego typu trwają przeciętnie od kilku do kilkunastu sekund (w zależności od mocy komputera oraz wielkości listy imputowej). Dlatego można powiedzieć, iż jest to główna metoda ekscerpcji neologizmów – nowych jednostek języka. Nawet w przypadku kilkuminutowych analiz dla bardzo obszernych list danych (rzędu miliona – trzech milionów słów tekstowych) możemy jednoznacznie powiedzieć, iż mamy do czynienia w tym wypadku z uzyskaniem wyniku natychmiastowego. Nie istnieje zatem inna, szybsza, bardziej efektywna metoda odnajdywania, ekscerpcji neologizmów niż ta, którą w tym miejscu prezentujemy. Główną zatem istotą tych badań jest wyeliminowanie z list słów (słowa te – jak pamiętamy – były wypisywane automatycznie z tekstów wstępnie dobieranych do badań), które analizator jest w stanie poprawnie zanalizować. Z punktu widzenia monitorowania innowacji leksykalnych te formy nas nie interesują.

18 Por. inne przykładowe nierozpoznane jednostki: *antipestilential, antisepticising, breakfastcupful, cantankerousest, chuckleheadedness, collectaneomania, companionability, comprehendingly, conciliatoriness, counterquestioned, damnfoolishness, demonstratorship, demonstratorships, dessertspoonful, disacidification, domesticability, doorkeeperishly, dunderheadedness, exceptionalness, experimentification, fragmentariness, honourablesness, humanitarianized, hypersensitiveness, hypocoristically, immaterialistic, incalculableness, incognizability, incorruptibleness, inhospitableness, interpretership, interpreterships, irrationalistic, irreconciliably, irreligiousness, knickerbockered, luxuriousenesse, malconfirmation, mediumistically, metropolitanism, microsporangium, neuropathically, noncommittingly, northwestwardly, overconscientiousness, overseriousness, overthoroughness, pachydermatously, pantheistically, paragraphically, phosphorescently, pictographically, preantepenultimate, preconsideration, psychopathologist, pusillanimously, respectableness, rightmindedness, schoolmastership, sciagraphically, sequesteredness, slabberdegullion, slantingdicular, supersensualist, superserviceably, tenderheartedness, thalpotasimeter, transmographied, troublesomeness, unapprehensible, unbelievableness, uncomplainingness, uncomplimentarily, understandingness, undesirableness, undiscriminatingly, undistinguishably, unecclesiastical, unecclesiastically, unexceptionableness, unexceptionably, unfavourableness, unintermittedly, unintermittingness, unmistakableness, unmistakeableness, unpicturesqueness, unprejudicedness, unpremeditatedly, unreflectiveness, unresistingness, unsentimentalist, unsubstantialness, unsympathizingly, untractableness, warmheartedness.*

W podsumowaniu należy także powiedzieć o wadach przedstawianej metody. Naturalnie, w miejscach, w których uzyskujemy pewne udogodnienia (np. szybkość analizy), pojawiają się pewne mankamenty stosowania tej metody (np. niedokładność wyniku). Główną wadą podnoszoną w przypadku analizy morfologicznej jest ograniczenie słownika (tj. wewnętrznej bazy słów) analizatora. Np. dla analizatora języka estońskiego (analizator o nazwie ESTMORF; por. Kaalep 1997), badania wskazywały na około 98% pokrycia tekstu, co oznacza, iż estońskie tłumaczenie tekstu „1984" G. Orwella zostało przez analizator zanalizowane w 98 procentach. Pozostałe 2% analizator uznał za formy nierozpoznane. Problemem jest to, iż w znacznej mierze na formy nierozpoznane składały się nazwy własne[19]. Zatem gdyby uznać, iż udało się w ten sposób wyekscerpować słowa nierozpoznane, byłyby to głównie słowa, które nie należą do grupy apelatywnych neologizmów (a ta kategoria nas głównie interesuje), lecz są nazwami własnymi. Kolejnym problemem jest konieczność ponownej, manualnej analizy wyniku. Oznacza to, że wynik, jaki otrzymujemy w komputerowej analizie, należy jeszcze raz zinterpretować, oddzielając w ramach form nierozpoznanych formy błędne oraz neologizmy. Ten proces z kolei jest etapem najbardziej pracochłonnym i właśnie na tym etapie ekscerpcji materiału nie sposób wyeliminować kompetencji lingwistycznej, tj. nadzoru użytkownika języka. Ostatnim problemem związanym z automatyczną ekscerpcją słownictwa jest kwestia niejednoznaczności wyniku. Wiąże się ta kwestia z tym, że komputer, a ściślej – analizator morfologiczny – może zanalizować jedynie formę graficzną danego słowa. Problem ten ilustruje analiza trzech słów: *piec*, *ekskomunika* oraz *liceum*. W przypadku analizy słowa *piec*:

piec[Sm3]+[SG]+[NOM]
piec[Sm3]+[SG]+[ACC]
piec[Vndk]=pie+c[B]

można mówić o interpretacji rzeczownikowej (mianownik lub biernik) oraz o interpretacji czasownikowej (bezokolicznik). Zatem aby jednoznacznie rozstrzygnąć, jaka nastąpiła analiza graficzna, należy sięgnąć do szerszego kontekstu, obserwując otoczenie napisu *piec*. W powyższym przypadku zanotowaliśmy problem dwojakiej interpretacji ze względu na różnicę części mowy (rzeczownik vs. czasownik). Poniżej zilustrujemy przykład różnicy wynikającej z homografii dwu słów będących tą samą częścią mowy, jednak różniących się rodzajem gramatycznym. W analizach:

19 Por. Mikheev 2002.

ekskomunika[Sf]=ekskomuni+ka[SG]+[NOM]
eks[sub']+komunik[Sm3]=komuni+ka[SG]+[GEN]

widzimy, iż w pierwszym wypadku mamy do czynienia z formą żeńską *ekskomunika*, natomiast możliwa jest także interpretacja *ekskomunika* jako połączenie prefiksu *eks-* oraz rzeczownika *komunik* (w przypadku dopełniacza: [GEN] w liczbie pojedynczej [SG]). Zatem ponownie – dla jednoznacznego określenia, jaką jednostkę analizujemy (rzeczownik rodzaju męskiego czy żeńskiego) – musimy odwołać się szerszego kontekstu, czyli do tekstu, w którym pojawiło się to słowo. Ostatni przykład[20] ilustruje homografię form słowa *liceum* w różnych formach przypadkowych (w liczbie pojedynczej). Tego typu problemy przede wszystkim sprawiają trudności w możliwie precyzyjnym określaniu liczby form przypadkowych (dla określonych przypadków), tzn. w tym wypadku niemożliwe jest bez sięgania w kontekst (składniowy) jednoznaczne określenie, w jakim przypadku gramatycznym pojawiło się dane słowo w tekście.

Prócz problemów natury gramatycznej, wynikających z niejednoznaczności interpretacji gramatycznej, w procesie automatycznej analizy pojawiają się problemy związane z interpretacją semantyczną (leksykalną) słów.

Ponownie należy podkreślić: program komputerowy dysponuje jedynie postacią graficzną danego słowa. Oznacza to, że niedostępne są w procesie analizy graficznej żadne inne informacje, poza informacjami właśnie graficznymi. Aby unaocznić skalę problemu, poddano analizie materiał z wybranych czasopism (z lat 1993 – 2003); celem analizy było ilościowe określenie nasycenia czasopisma nazwami zwierząt. Okazało się, że w roku 1999 nienaturalnie (w stosunku do innych lat) wysoko na liście rangowej zwierząt uplasowała się *pluskwa*. W roku zaś 2003 – *lew*. Te „przekłamania" związane są z dwoma faktami pozajęzykowymi. Otóż w roku 1999 w tekstach bardzo często pojawiała się fraza: *pluskwa milenijna* związana z komputerowym problemem zapisu daty roku 2000. Na wysoką frekwencję słowa *lew* naturalnie wpłynęła tzw. *afera Rywina*, w kontekście której imię *Lew* pojawiało się bardzo często. Jak widzimy zatem, analiza komputerowa, czyli automatyczna analiza morfologiczna prowadząca do

20 Analysis of "liceum":
liceum[Sn]=lice+um[SG]+[NOM]
liceum[Sn]=lice+um[SG]+[GEN]
liceum[Sn]=lice+um[SG]+[DAT]
liceum[Sn]=lice+um[SG]+[ACC]
liceum[Sn]=lice+um[SG]+[INS]
liceum[Sn]=lice+um[SG]+[LOC]
liceum[Sn]=lice+um[SG]+[VOC]

ekscerpcji neologizmów[21] jest czynnością bardzo „niebezpieczną", tj. następować po niej musi odpowiednia analiza manualna, dokonywana przez kompetentnego lingwistę[22], tj. analizy tej – analizy dokonanej przez użytkownika języka – nie sposób wyeliminować na żadnym etapie poszukiwania nowych jednostek języka lub też określania ich frekwencji.

Bibliografia

Aarts, J., Meijs, W. (red.). 1984. *Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi Bv Editions.

Bańko, M. 2001. *Z pogranicza leksykografii i językoznawstwa. Studia o słowniku jednojęzycznym*. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego.

Bańko, M. 2002. *Wykłady z polskiej fleksji*. Warszawa: PWN.

Baroni, M., Matiasek, J., Trost, H. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*, 48 – 57.

Bień, J.S., Szafran, K. 1996. Analiza języka polskiego w Instytucie Informatyki UW. W: Vetulani, Z. et al. (red.). 1996, 77 – 81.

Bień, J.S., Szafran, K. 2001. Analiza morfologiczna języka polskiego w praktyce. *Biuletyn Polskiego Towarzystwa Językoznawczego* LVII, 171 – 184.

21 Innymi sposobami ekscerpcji neologizmów wykorzystywanymi przez nas w codziennej pracy językoznawczej są:
a) ekscerpcja użyć tekstowych poprzedzonych jednostkami typu: *tzw., tak zwane, nazywa się, określa się, określa się mianem* itp.: „Przed Sądem Rejonowym w Dąbrowie Tarnowskiej rozpoczął się proces 10 dębickich policjantów, w tym byłych szefów komendy rejonowej, z komendantem Leszkiem S. To efekt tzw. **afery notatkowej**. Prokuratura zarzuca im, że w statystykach nie wykazywali wszystkich zgłoszeń o popełnianych przestępstwach, co zawyżało wskaźniki wykrywalności" (21 numer Obserwatora Lokalnego; Dębicki Portal).
b) ekscerpcja jednostek w cudzysłowach:
„W kontekście «wyczynów» sprzed lat niektórych policjantów z Dębicy znanych bardziej pod nazwą **«afery notatkowej»** wydarzenia ostatnich tygodni jeszcze bardziej zaskakują. Sądowe procesy widać niewiele ich nauczyły".
Obok użyć bezcudzysłowowych: „«Nie dostrzegali, że bogactwo osobiste przekłada się na dobrobyt społeczny» – dziwiła się w jedynym, udzielonym po wybuchu **afery notatkowej**, wywiadzie dla telewizyjnej «Summy zdarzeń»" (Polityka 47/2004).

22 Por. efekty ekscerpcyjne uzyskane w pracy Wawrzyńczyk 2000.

Buttler, D. 1962. Neologizm i terminy pokrewne. *Poradnik Językowy* 5 – 6, 235 – 244

Čermák, F. 2001. Substance of Idioms: perennial problems, lack of data or theory? *International Journal of Lexicography* 14, 1 – 20.

Daciuk, J. 1999. A Module for Treatment of Unknown Words. *Speech and Language Technology* 3, 165 – 169.

Degand, L., Bestgen, Y. 2003. Towards Automatic Retrieval of Idioms in French Newspaper Corpora. *Literary and Linguistic Computing* 18 (3), 249 – 259.

Gajda, S. (red.). 2001. *Język polski*. Opole: UO, IFP.

Gajda, S. (red.). 2003. *Językoznawstwo w Polsce. Stan i perspektywy*. Opole: UO.

Gilquin, G. 2002. Automatic retrieval of syntactic structures: The quest for the Holy Grail. *International Journal of Corpus Linguistics* 7 (2), 183 – 214.

Goldsmith, J. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27, 153 – 198.

Graliński, F. 1998. Realizacja półautomatycznej ekstrakcji leksemów występujących w korpusie polskich tekstów informatycznych. *Speech and Language Technology* 2, 127 – 135.

Graliński, F., Krynicki, G. 2000. Word-Formation Analysis in Polish-to-English Machine Translation. *Speech and Language Technology* 4, 185 – 203.

Han, Chung-hye, Han, Na-Rae, Ko, Eon-Suk, Palmer, M. 2002. Development and Evaluation of a Korean Treebank and its Application to NLP. *Language and Information* 6 (1), 123 – 138.

Jadacka, H. 2001. *System słowotwórczy polszczyzny (1945 – 2000)*. Warszawa: PWN.

Jones, S., Sinclair, J.M. 1974. English lexical collocations. *Cahiers de Lexicologie* 24, 15 – 61.

Kaalep, H-J. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. *Computers and the Humanities* 31, 115 – 133.

Kang, Seung-Shik, Kim, Yung Taek. 1994. Syllable-based Model for the Korean Morphology. *COLING 1994: The 15th International Conference on Computational Linguistics* 1, 221 – 226.

Kiefer, F., Kiss, G., Pajzs, J. (red.). 1992. *Papers in Computational Lexicography. Complex-92*. Budapest: Hungarian Academy of Sciences, Linguistics Institute.

Koval, S., Beliaeva, L., Kogan, L., Mikhailov, A., Nilolaev, V., Piotrowski, R., Tomach, Y. 2000. Morphological representation in PC-based text processing systems. *Literary and Linguistic Computing* 15 (2), 131 – 156.

Kučera, K. 2002. The Czech National Corpus: Principles, Design, and Results. *Literary and Linguistic Computing* 17 (2), 245 – 257.

Kuraszkiewicz, W. 1973. Obfitość słownictwa w kilku dużych tekstach polskich. *Studia Polonistyczne* I, 45 – 63.

Kurcz, I. 1973. Założenia ogólne frekwencyjnych badań nad słownictwem i cele, którym te badania służą. *Przegląd Pedagogiczny* 1 – 2, 65 – 75.

Kurkowska, H. 1956. *O nowym słownictwie polskim*. Warszawa: Wiedza Powszechna.

Kwon, Oh-Wook, Park, Jun. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication* 39, 287 – 300.

Lee, Do-Gil, Rim, Hae-Chang, Lim, Heui-Seok. 2003. A Syllable Based Word Recognition Model for Korean Noun Extraction. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 471 – 478.

Lee, Gary Geunbae, Cha, Jeongwon, Lee, Jong-Hyeok. 2002. Syllable-Pattern-Based Unknown-Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean. *Computational Linguistics* 28 (1), 53 – 70.

Lubaś, W., Sowa, F. (red.). 1993. *Wokół słownika współczesnego języka polskiego. III. Zakres selekcji i informacji*. Kraków: IJP PAN.

Mańczak, W. 1959. Pojęcie ilości w języku. *Studia Filozoficzne* 60, 111 – 127.

Mikheev, A. 2002. Periods, Capitalized Words, etc. *Computational Linguistics* 28 (3), 289 – 318.

Nagórko-Kufel, A. 1978. Statystyczna struktura słownictwa motywowanego. *Poradnik Językowy* 3, 99 – 105.

Obrębski, T. 1998. Wykorzystanie lematyzatora słownika POLEX do oznaczania form wyrazowych w korpusie tekstów informatycznych. *Speech and Language Technology* 2, 113 – 126.

Oflazer, K. 1996. Error-tolerant finite state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics* 22 (1), 73 – 89.

Park, Bong-Rae, Hwang, Young-Sook, Rim, Hae-Chang. 1997. Recognizing Korean Unknown Words by Comparatively Analyzing Example Words. *Proceedings of the 1997 International Conference on Computer Processing of Oriental Languages*, 127 – 132.

Park, Bong-Rae, Hwang, Young-Sook, Rim, Hae-Chang. 1998. Eliminating Implausible Korean Morphological Interpretations by Using History of Previous Analysis and Lexical Association. *Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence*, 31 – 36.

Pawłowski, A. 2003. Uwagi na temat korpusu języka polskiego (reprezentatywność, aktualność, nazwa). W: Gajda, S. (red.). 2003, 162 – 180.

Proszeky, G., Tihanyi, L. 1992. A Fast Morphological Analyzer for Lemmatizing Agglutinative Languages. W: Kiefer, F. et al. (red.). 1992, 265 – 278

Saloni, Z. 1992. Co istnieje, a co nie istnieje we fleksji polskiej. *Prace Filologiczne* XXXVII, 75 – 87.

Sambor, J. 1975. *O słownictwie statystycznie rzadkim (na materiale derywatów we współczesnej publicystyce polskiej).* Warszawa: PWN.

Sgarbas, K.N., Fakotakis, N.K., Kokkinakis, G.K. 1998. A PC-KIMMO-based Bi-directional Graphemic/phonetic Converter for Modern Greek. *Literary and Linguistic Computing* 13 (2), 65 – 76.

Sgarbas, K.N., Londos, G.E., Fakotakis, N.D., Kokkinakis. G.K. 2003. The WATCHER Project: Building an Agent for Automatic Extraction of Language Resources from the Internet. *Literary and Linguistic Computing* 18, 449 – 464.

Sharma, U., Kalita, J., Das, R. 2002. Unsupervised Learning of Morphology for Building Lexicon for a Highly Inflectional Language. *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/ SIGPHON-2002*, 1 – 10.

Sinclair, J.M. 1992. The automatic analysis of corpora. W: Svartvik, J. (red.). 1992, 379 – 397.

Smith, G.W. 1991. *Computers and Human Language*. New York – Oxford: Oxford University Press.

Smółkowa, T. 2001. *Neologizmy we współczesnej leksyce polskiej*. Kraków: IJP PAN.

Solak, A., Oflazer, K. 1993. Design and Implementation of a Spelling Checker for Turkish. *Literary and Linguistic Computing* 8 (3), 113 – 130.

Stoberski, Z. 1976. O centralną rejestrację neologizmów naukowych. *Poradnik Językowy* 4, 186 – 189.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7 (2), 215 – 244.

Suszczańska, N., Forczek, M., Migas, A. 2000. Wieloetapowy analizator morfologiczny. *Speech and Language Technology* 4, 55 – 65.

Svartvik, J. (red.). 1992. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm, 4 – 8 August 1991*. Berlin: Mouton de Gruyter.

Szafran, K. 1993. *Automatyczna analiza fleksyjna tekstu polskiego (na podstawie Schematycznego indeksu a tergo Jana Tokarskiego)*. Uniwersytet Warszawski. Wydział Polonistyki. Niepublikowana praca doktorska.

Tambouratzis, G., Carayannis, G. 2001. Automatic Corpora-based Stemming in Greek. *Literary and Linguistic Computing* 16 (4), 445 – 466.

Teahan, W.J., Wen, Y., McNab, R., Witten, I.H. 2000. A Compression-based Algorithm for Chinese Word Segmentation. *Computational Linguistics* 26 (3), 375 – 393.

Waszakowa, K. 1997. Rola kontekstu i sytuacji w rozumieniu neologizmów. *Biuletyn Polskiego Towarzystwa Językoznawczego* 53, 121 – 132.

Waszakowa, K. 2001. System słowotwórczy. W: Gajda, S. (red.). 2001, 88 – 107.

Wawrzyńczyk, J. 1992. *Chronologizacja słownictwa nowopolskiego. W poszukiwaniu źródeł dokumentacyjnych neologizmów powojennych*. Toruń: UMK.

Wawrzyńczyk, J. 1993. Uwagi o rejestracji neologizmów polszczyzny dwudziestowiecznej. W: Lubaś, W., Sowa, F. (red.). 1993, 33 – 40.

Wawrzyńczyk, J. 2000. *Słownik bibliograficzny języka polskiego: wersja przedelektroniczna*. T.1, A-Ć. Warszawa: Warszawa: Instytut Informacji Naukowej i Studiów Bibliologicznych Uniwersytetu Warszawskiego.

Wołosz, R. 2000. *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*. Uniwersytet Warszawski. Wydział Polonistyki. Niepublikowana rozprawa doktorska.

Wypych, M. 2002. Stochastic Spelling Correction of Texts in Polish. *Speech and Language Technology* 6, 243 – 250.

Vetulani, Z., Abramowicz, W., Vetulani, G. (red.). 1996. *Język i technologia.* Warszawa: Akademicka Oficyna Wydawnicza PLJ.

Vetulani, Z., Walczak, B., Obrębski, T., Vetulani, G. 1998. *Jednoznaczne kodowanie fleksji rzeczownika i jego zastosowanie w słownikach elektronicznych – format POLEX.* Poznań: Wydawnictwo Naukowe UAM.

Yamamoto, M., Church, K.W. 2001. Using Suffix Arrays to Compute Term Frequency and Document Frequency for all Substrings in a Corpus. *Computational Linguistics* 27 (1), 1 – 30.

Yoon, J. 2001. Efficient dependency analysis for Korean sentences based on lexical association and multi-layered chunking. *Literary and Linguistic Computing* 16 (3), 265 – 285.

Zagrodnikowa, A. 1982. *Nowe wyrazy i wyrażenia w prasie*. Kraków: OBP.

Yoon, J. 2001. Efficient dependency analysis for Korean sentences based on lexical association and multi-layered chunking. *Literary and Linguistic Computing* 16 (3), 265 – 285.

Zagrodnikowa, A. 1982. *Nowe wyrazy i wyrażenia w prasie*. Kraków: OBP.