Investigationes Linguisticae, vol. XII, Poznań, December 2005

Automatyczna analiza morfologiczna jako narzędzie oceny poprawności wyrazów języka rosyjskiego

Automated morphological analysis as a assessment tool for correctness of Russian words

Jurij Fiedoruszkow

Institute of Linguistics, Adam Mickiewicz University
ul. Międzychodzka 5, 60-371 Poznań, POLAND

jerfed@wp.pl

Abstract

Processes of lemmatizing are beginning to play a very important role in contemporary linguistic methods. A lemmatizer as an instrument in computer analysis morphologically analizes texts and corpuses of texts and at the same time functions as a program module and a lemmatizing algorithm. The identification of lexemes depends on the organization of the dictionary data. One of the examples of the practical results of lemmatization is the set of unrecognized words (with a precise determination of the membership of morphological markers and parts of speech) in a language. In this paper, the methodological description of the use of the program Lemmatizer with the use of an electronic text (as an example) in Russian is proposed.

1. Wstęp

Analiza morfologiczna wyrazów badanego tekstu może być celem samym w sobie, oferując wysokie wartości poznawcze (Bień, Szafran 2001, Rabiega-Wiśniewska, Rudolf 2002). W dalszej kolejności analiza ta ma na celu oznaczenie wartości cech morfologicznych na potrzeby analizy dalszej, np. analizy syntaktycznej (Vetulani 2004). Analizator jest jednocześnie modułem programowym oraz algorytmem *lematyzacji* polegającej na identyfikacji (Szafran 1997) – jednoznacznej lub nie – leksemu. Identyfikacja ta wynika z organizacji danych słownikowych, które stanowią integralną część każdego analizatora morfologicznego. Charakter takiej identyfikacji pozwala na określenie przynależności cech morfologicznych (m.in. przypadek gramatyczny, liczba, aspekt itp.) do określonego leksemu oraz odpowiedniej części mowy (kategorii morfo-syntaktycznej). Ustalenie takiej przynależności oraz określenie formy kanonicznej, hasłowej nosi nazwę *lematyzacji*.

Analizator morfologiczny to program, który zwraca (dla tekstu wejściowego) listę jednostek wejściowych razem z adekwatnym do tego wyniku analizy opisem gramatycznym. Dlatego często lematyzatory używane są w celu sprawdzania pisowni. Lematyzator pojmowany w wąskim

zakresie to analizator morfologiczny sprowadzający formę tekstową wyrazu do formy podstawowej, zlematyzowanej (nazywa się ją także lematem[1]). W niniejszej pracy interesuje nas nieco inny aspekt wykorzystania analizatora: obserwacja wyników w postaci listy nierozpoznanych przez lematyzator słów. Pracę taką dla języka rosyjskiego przedstawił m.in. Wierzchoń (2005). Identyfikacja leksemu pozwala na dalszą analizę – dotarcie do informacji zawartej w słownikach, np. gramatycznych (np. Зализняк 1987).

2. Program LEMMATIZER

LEMMATIZER został stworzony przez grupę moskiewskich lingwistów, zajmujących się projektem *Dialing*[2]. Projekt dotyczy automatycznej analizy tekstu języka rosyjskiego, niemieckiego i angielskiego. Analiza m.in. jest ściśle powiązana z tłumaczeniem tekstów w danych językach, z ich składnią i morfologią, semantyką powierzchniową, transformacją tekstów. LEMMATIZER (dalej zwany *Programem*) udostępniony jest jako wersja demonstracyjna na serwerze ogólnodostępnym.

Program w różnych odmianach funkcjonuje na platformie Win2000/NT/XP/ME, Linux. Pakiet programu zawiera rodzaj rosyjskiej morfologii, słowniki binarne języka rosyjskiego, angielskiego oraz niemieckiego (biblioteki). Do pracy analizatora potrzebny jest coclass dla języka rosyjskiego zwany LemmatizerRussian-.

Program jest przygotowany do analizy morfologicznej, tzn. do wspomnianej wyżej identyfikacji oraz do budowania interpretacji (por. APPENDIX; pkt. 1.) morfologicznej słów tekstu wejściowego i – jako do stanu przejściowego – do segmentacji dowolnego zdania w języku rosyjskim.

3. Materiał badawczy

Wybrany tekst wejściowy do analizatora dotyczy słownictwa (terminologii) z dziedziny księgowości. Большой бухгалтерский словарь (Азрилиян 1999) – Wielki słownik księgowości – pod red. A. Azriliana, ma również wersję elektroniczną (dalej: e-wersja), która jest w całości udostępniona[3] na rosyjskim serwerze[4]. To głównie ta dostępność skłoniła nas do wyboru tego tekstu. Uważa się między innymi, że, po pierwsze, zgodnie z wynikami badań rosyjskiego czasopisma Мир ПК zbiór elektroniczny słowników LingVo firmy ABBYY jest liderem: wybiera go około 92 % użytkowników słowników kategorii Электронные словари. Po drugie, przy przygotowaniu słownika Большой бухгалтерский словарь użyty został elektroniczny uniwersalny system prawny ВАШЕ ПРАВО 98.

4. E-wersja – charakterystyka

Dane elektroniczne tekstu wejściowego. Właściwości e-wersji Słownika:
rozmiar: 7,87 MB;
liczba stron: 1268;
liczba znaków: 3.311.249;
liczba wyrazów: 463173;
wiersze: 63039;
akapity: 8643;
znaki ze spacjami: 3.766.273.

---

[1] Notujemy różne definicje *lematu*, m.in.:
1. Lemat (w sensie logicznym) – to twierdzenie pomocnicze, służące do udowodnienia innego twierdzenia, bardziej zasadniczego w tym momencie (Por. Wojnowski 2003: 442).
2. Lemma (w sensie lingwistycznym) – lemat, wyraz hasłowy hasło (Por. Linde-Usiekniewicz 2002: 676).

[2] www.aot.ru

[3] Kwestia związana z zasadami umieszczania słowników w Internecie omówiona została m.in. w dokumencie: http://www.langust.ru/news/01_07_02.shtml.

[4] http://www.stavropolaudit.ru/news.html

5. Przykłady

Przykłady tekstu wejściowego. Hasła słownikowe:

(…)
"АВАНСЫ ПОСТАВЩИКАМ" – название активного счета, отражающего денежные средства, уплаченные авансом до получения (поставки) товаров или услуг.

АВАРИЯ – 1. ущерб и убытки, причиненные транспортному средству, грузу и фрахту в процессе перевозки. В зависимости от характера и принципов распределения убытков между участниками морской перевозки авария подразделяется на общую аварию и частную аварию; 2. в страховании от пожара является уменьшением выплаты страхового возмещения вследствии недострахования.

АВЕРАЖ – среднее количество товара.
(…)

6. Analiza wyjściowa

Parametry analizy związane są z danymi poszczególnego elementu tekstu. Jako elementy wyjściowe interesują nas nierozpoznane przez *Program* wyrazy złożone i proste oraz kategorie morfologiczne słów na bazie *gramemów* i *lematów* [por. APPENDIX; pkt. 2.]. Dlatego wykonujemy kroki eliminacji niektórych elementów analizy:

krok 1
Eliminacja znaków interpunkcyjnych, skrótów.

krok 2
Analiza morfologiczna. Każdy wers analizy słowoformy oznacza zbiór wartości kategorii gramatycznych danego słowa za pomocą kodu oraz weryfikacją pozytywną:

части 4560 5 RLE aa +Фа ЧАСТЬ гбгвгегжгй 112364 0
части 4560 5 RLE aa +Уо ЧАСТИТЬ кл 55155 0

lub negatywną:

алфавитно-гнездовому 4267 20 RLE aa -?? АЛФАВИТНЫЙ-ГНЕЗДОВОЙ йвйо -1 0
алфавитно-гнездовому 4267 20 RLE aa -?? АЛФАВИТНЫЙ-ГНЕЗДОВЫЙ йвйо -1 0

Pierwszy przykład pokazuje, że *Program* rozpoznaje słowo *части* oraz podaje dwie możliwe lematy: wyraz *часть* oraz *частить*, uznając w ten sposób, że słowoforma wejściowa może być rzeczownikiem bądź formą rozkazującą czasownika *частить*. Innymi słowy, jedno słowo posiada kilka interpretacji morfologicznych. W drugim przykładzie *Program* także podaje potencjalnie możliwe lematy, lecz nie rozpoznaje leksemu względem słownika (Азрилиян 1999). Tabele kodów dla analizy morfologicznej znajdują się w załącznikach *Programu* i udostępnione są na wyżej wspomnianym serwerze. Poniżej podane są przykładowe kody dla rzeczownika rosyjskiego *стол* (fragment):

аа	С	мр,но	ед,им	стол
аб	С	мр,но	ед,рд	стола
ав	С	мр,но	ед,дт	столу
аг	С	мр,но	ед,вн	стол
ад	С	мр,но	ед,тв	столом
ае	С	мр,но	ед,пр	столе
аж	С	мр,но	мн,им	столы
аз	С	мр,но	мн,рд	столов
аи	С	мр,но	мн,дт	столам
ай	С	мр,но	мн,вн	столы
ак	С	мр,но	мн,тв	столами
ал	С	мр,но	мн,пр	столах

Elementy pierwszej kolumny używane są w charakterystyce morfologicznej każdej części mowy w analizie wyjściowej *Programu*.

krok 3
Dany krok dotyczy eliminacji wszystkich pozytywnie sprawdzonych słów w uzyskanej analizie, na przykład:

долгосрочного, исполнительной, заинтересованными, заинтересованных, пропорционально, пропорциональном, профессиональный, представительное, положительная, перечисленных, краткосрочные, определенного, определенному, непосредственно, непосредственного, ответственным, ответственных, самостоятельно, самостоятельного, самостоятельном, самостоятельными, хозяйственного, хозяйственной, хозяйственные, хозяйственных, централизованных, функциональной itd.

Czynność ta pozwala na wyróżnienie wszystkich nieznanych przez słowniki *Programu* słów, np.:
(…)

терминах-словосочетаниях	4362 24 RLE аа -?? ТЕРМИН-СЛОВОСОЧЕТАНИЕ ел -1 0
внегнездовых	5534 12 RLE аа -?? ВНЕГНЕЗДОВЫЙ йуйхйч -1 0
внегнездовых	5534 12 RLE аа -?? ВНЕГНЕЗДОВОЙ йуйхйч -1 0
одностороннем	8213 13 RLE аа -?? ОДНОСТОРОННИЙ йейс -1 0
КОМПАНИЯМ-ФИЛИАЛАМ	11402 18 RLE АА -?? КОМПАНИЯ-ФИЛИАЛ аи -1 0
оперативно-техническому	20020 23 RLE аа -?? ОПЕРАТИВНЫЙ-ТЕХНИЧЕСКИЙ йвйо -1 0
оперативно-техническому	20020 23 RLE аа -?? ОПЕРАТИВНО-ТЕХНИЧЕСКИЙ йвйо -1 0
оперативно-техническому	20020 23 RLE аа -?? ОПЕРАТИВНО-ТЕХНИЧЕСКИЙ йвйо -1 0
банках-корреспондентах	21322 22 RLE аа -?? БАНК-КОРРЕСПОНДЕНТ ал -1 0
банках-корреспондентах	21322 22 RLE аа -?? БАНКА-КОРРЕСПОНДЕНТ ал -1 0
АГЕНТ-ОПТОВИК	26624 13 RLE АА SENT1 -?? АГЕНТ-ОПТОВИК аа -1 0
АГЕНТ-ОПТОВИК	26624 13 RLE АА SENT1 -?? АГЕНТ-ОПТОВИК аа -1 0
АГЕНТ-ТРАНСФЕР	26823 14 RLE АА SENT1 -?? АГЕНТ-ТРАНСФЕР аааг -1 0
АГЕНТ-ТРАНСФЕР	6823 14 RLE АА SENT1 -?? АГЕНТ-ТРАНСФЕР аааг -1 0
Трансфер-агент	27300 14 RLE Аа NAM? SENT1 -?? ТРАНСФЕР-АГЕНТ аа -1 0
Трансфер-агент	27300 14 RLE Аа NAM? SENT1 -?? ТРАНСФЕР-АГЕНТ аааг -1 0
А-ДАТО	30184 6 RLE АА SENT1 -?? А-ДАТО яа -1 0
а-дато	30316 6 RLE аа -?? А-ДАТО яа -1 0
учетно-вычислительные	32828 21 RLE аа -?? УЧЕТНЫЙ-ВЫЧИСЛИТЕЛЬНЫЙ йтРь -1 0
Банк-эмитент	33950 12 RLE Аа NAM? SENT1 -?? БАНК-ЭМИТЕНТ аа -1 0
банка-эмитента	34760 14 RLE аа -?? БАНК-ЭМИТЕНТ абаг -1 0
банка-эмитента	34760 14 RLE аа -?? БАНКА-ЭМИТЕНТ абаг -1 0
банка-эмитента	36100 14 RLE аа -?? БАНК-ЭМИТЕНТ абаг -1 0
банка-эмитента	36100 14 RLE аа -?? БАНКА-ЭМИТЕНТ абаг -1 0
ное	38774 3 RLE аа -?? НЫЙ ймйп -1 0
Одностороннее	40945 13 RLE Аа NAM? SENT1 -?? ОДНОСТОРОННИЙ ймйпйю -10

односторонней	42224 13 RLE aa -?? ОДНОСТОРОННИЙ йзйийкйлйю -1 0
банк-эмитент	42742 12 RLE aa -?? БАНК-ЭМИТЕНТ aa -1 0
предприятий-экспортеров	44353 23 RLE aa -?? ПРЕДПРИЯТИЕ-ЭКСПОРТЕР азай -1 0
АКТ-ИЗВЕЩЕНИЕ	48593 13 RLE AA SENT1 -?? АКТ-ИЗВЕЩЕНИЕ еаег -1 0
ПРИЕМКИ-ПЕРЕДАЧИ	55620 16 RLE AA -?? ПРИЕМКА-ПЕРЕДАЧА гбгжгй -1 0
акте-требовании	59888 15 RLE aa -?? АКТ-ТРЕБОВАНИЕ ее -1 0
акте-требовании	59967 15 RLE aa -?? АКТ-ТРЕБОВАНИЕ ее -1 0
акты-требования	60043 15 RLE aa -?? АКТ-ТРЕБОВАНИЕ ебежей -1 0

Najczęściej są to niespotykane w ogólnych słownikach rosyjskich słowa (np. terminy). Przykłady tych wyrażeń znajdziemy w analizowanym materiale. Na przykład potwierdzenie istnienia jednostki „а-дато”:

а-дато 30316 6 RLE aa -?? А-ДАТО яа -1 0

szukamy w źródle:

А-ДАТО – число, от которого дан тот или иной документ, например, вексель дан от такого-то числа, через столько-то времени заплатить а-дато.

Program podaje w niektórych wypadkach kilka interpretacji, świadczących o przynależności nieznanego słowa do jednej formy kanonicznej (lemm), którą proponuje:

алфавитно-гнездовому 4267 20 RLE aa -?? АЛФАВИТНЫЙ-ГНЕЗДОВОЙ йвйо -1 0
алфавитно-гнездовому 4267 20 RLE aa -?? АЛФАВИТНЫЙ-ГНЕЗДОВЫЙ йвйо -1 0

krok 4

Po usunięciu wszystkich prawych części analizy (tj. informacji morfologicznych i form zlematyzowanych) otrzymujemy zbiór w postaci 2844 powtarzających się jednostek językowych, np.:

А-ДАТО	амортизируемого	национально-административной
А-МЕТА	амортизируемых	ное
АГЕНТ-ОПТОВИК	аналитико-методический	нормативно-справочного
АГЕНТ-ТРАНСФЕР	англо-говорящих	нормативно-справочной
АКТ-ИЗВЕЩЕНИЕ	багажно-грузовых	овой
АКТ-ТРЕБОВАНИЕ	баланс-брутто	одностороннее
БАЛАНС-ЭКСТЕРН	баланс-нетто	односторонней
БАНК-АГЕНТ	баланс-это	одностороннем
БАНК-АКЦЕПТАНТ	банк-корреспондент	однотипны
БАНК-ГАРАНТ	банк-эмитент	оперативно-бухгалтерскому
БАНК-КОРРЕСПОНДЕНТ	банка-акцептанта	оперативно-техническому
БАНК-МОСТ	банка-банкрота	организационно-прав
БАНК-РЕМИТЕНТ	банка-заемщика	организационно-правовая
БАР-КОД	банка-эмитента	организационно-правовой
БЕЗНОМИНАЛЬНЫЕ	банках-корреспондентах	организационно-правовую
БИЗНЕС-ОПЕРАЦИЯ	банков-корреспондентов	организационно-распорядительную
БИЗНЕС-СДЕЛКА	банком-плательщиком	отчетно-статистическую
БИЗНЕС-ШКОЛА	банком-ремитентом	показатели-индикаторы
БЛАНК-ВАУЧЕР	банку-корреспонденту	полубрутто
БРУТТО-АРЕНДА	бизнес-деятельности	полугодовые

БРУТТО-БАЛАНС
БРУТТО-ДОЛЯ
БРУТТО-ДОХОД
БРУТТО-ПРЕМИЯ
БРУТТО-ПРОДАЖИ
БРУТТО-ПРОЦЕНТЫ
БРУТТО-ТОННАЖ
БУХГАЛТЕР-АУДИТОР
БУХГАЛТЕР-РЕВИЗОР
БЮДЖЕТ-БРУТТО
Банк-эмитент
Бирда
Блатова
Бреттон-Вулского
ВАУЧЕР-ЧЕК
ВЕКСЕЛЬ-ПРИМА
ВЕКСЕЛЬ-СЕКУНДА
ВЕКСЕЛЬ-ТЕРЦИЯ
ГРУППИРОВОЧНАЯ
ДЕГРЕССИВНАЯ
МИНЕРАЛЬНО-СЫРЬЕВОЙ
НАЛОГООБ-ЛАГАЕМАЯ
ОСНОВНОЙ
Одностороннее
ПРЕПОРУЧИТЕЛЬНЫЕ
ПРИЕМКИ-ПЕРЕДАЧИ
ПРИЕМКИ-СДАЧИ
ПРИРОСТНОЙ
ПРОИЗВОДСТВЕННО-СБЫТОВОЙ
РАСЧЕТНО-ПЛАТЕЖНАЯ
Расчетно-платежная
Расчетно-платежные
СВОБОДНАЯЗАГРУЗОЧНАЯ
СПРОСА-ПРЕДЛОЖЕНИЯ
ССУДНО-СБЕРЕГАТЕЛЬНАЯ
СТАТИКО-ДИНАМИЧЕСКОГО
ТАРА-ТАРИФУ
Трансфер-агент
ФИНАНСОВО-ПЛАНОВЫЙ
ФУНКЦИОНАЛЬНО-СТОИМОСТНОЙ
Франческо
ШТРИХ-КОД
а-дато
административно-хозяйственные
административно-хозяйственных
акте-требовании
акты-требования
акциями-барометрами

бизнес-операций
бланка-ваучера
брутто-ставки
бухгалтер-практик
бухгалтеров-аналитиков
бухгалтерско-аудиторских
валютно-кредитная
валютно-кредитных
валютно-финансовым
ведущегося
ведущуюся
вексель-документ
внегнездовых
гильоширных
гонорар-вознаграждение
горно-химического
государства-заемщика
группировочных
двустороннюю
двухсторонняя
двухуровневую
двухцветных
документографической
жилищно-гражданского
журналам-ордерам
журнале-ордере
журналов-ордеров
журнально-ордерная
журнально-ордерной
заем-шик
импортер-векселедатель
институалистического
информационно-ориентирующая
информационно-справочная
информационно-справочного
компанией-эмитентом
компаний-конкурентов
контр-брокера
корпорации-эмитента
культурно-бытового
купле-продаже
купли-продажи
куплю-продажу
купля-продажа
лицо-акцептант
математико-статистические
метаинформационной
минерально-сырьевой

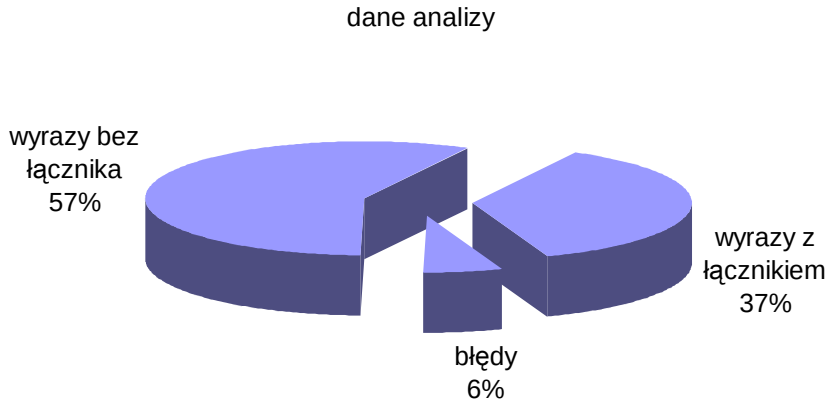
полугосударственные
полунетто
почвенно-климатических
предприятием-покупателем
предприятий-экспортеров
предприятия-должника
предприятия-поставщики
приростных
программно-инструментальное
производственно-технологические
разрезелицевых
расчетно-кредитного
расчетно-платежная
расчетно-платежной
расчетно-платежную
расчетно-платежным
расчетно-платежных
санаторно-курортное
сводно-группировочных
спирта-сырца
статико-динамического
сток-брокером
стран-заемщиков
стран-членов
странами-участницами
страны-члены
счет-счету
счетов-фактур
терминах-словосочетаниях
товарно-сопроводительных
товаро-материальными
товаропередаточным
транспортно-заготовительных
учетно-вычислительные
учетно-вычислительных
финансово-денежных
финансово-хозяйственная
финансово-хозяйственной
финансово-хозяйственных
финансово-экономического
хозяйственно-операционные
хозяйственно-финансовой
хозяйственно-финансовых
членов-учредителей
членом-корреспондентом
эксперт-оценщик
экспертно-аналитическую
являетсяосновным

алфавитно-гнездовому	натурально-стоимостным
алфавитно-предметный	натурально-стоимостнымнатурально-стоимостным

itd.

krok 5

Po eliminacji wszystkich powtórzeń otrzymujemy 1750 niepowtarzalnych wyrazów. Wyrazy te podzielono w sposób następujący:



Wyrazy z łącznikiem

Uzyskane w badaniu wyrazy z łącznikiem to np.:

верблюды-производители,	сводно-группировочных
вещественно-натуральной,	спирта-сырца
почвенно-климатических	статико-динамического
предприятием-покупателем	сток-брокером
предприятий-экспортеров	стран-заемщиков
предприятия-должника	стран-членов
предприятия-поставщики	странами-участницами
программно-инструментальное	страны-члены
производственно-технологические	счет-счету
расчетно-кредитного	счетов-фактур
санаторно-курортное	терминах-словосочетаниях

Wyrazy te stanowią najciekawszą grupę ze względu na specyfikę ich ujęcia przez *Program*. Przykłady:

a) банка-эмитента	35212 14 RLE aa -?? БАНК-ЭМИТЕНТ абаг -1 0
банка-эмитента	35212 14 RLE aa -?? БАНКА-ЭМИТЕНТ абаг -1 0
b) бухгалтеры-практики	763204 19 RLE aa -?? БУХГАЛТЕР-ПРАКТИКА гбгжгй -1 0
бухгалтеры-практики	763204 19 RLE aa -?? БУХГАЛТЕР-ПРАКТИК аж -1 0

wskazują na dwie możliwe (*Nom, Sing*) formy (w danym przypadku jest to kategoria Rodzaju) pierwszej części połączenia *банк-эмитент* (przykład *a*) oraz drugiej części połączenia *бухгалтеры-практики* (przykład *b*).

Po eliminacji wszelkich powtórzeń, wynikających z różnych form paradygmatu, stanowią one 37% słów (bez uwzględnienia wielkości liter). Np.:

журнала-ордера	773346 14 RLE aa -?? ЖУРНАЛ-ОРДЕР абажай -1 0
ЖУРНАЛЫ-ОРДЕРА	773435 14 RLE AA SENT1 -?? ЖУРНАЛ-ОРДЕР абажай -1 0

6.1. Błędy ortograficzne

Błędy ortograficzne wyróżnione zostały poprzez nieautomatyczną obserwację wyników w postaci słów nierozpoznanych przez *Program*. Na przykład:

оизводственно-экономический	БИХЕВИОРИ-СТИЧЕСКАЯ
АВТОРИЗАЦИЯКРЕДИТНОЙ	Взависимости
баланс-это	Включающийналог
бесполуфабрикатномварианте	ВЛАДЕЛЬЦЕВТРАНСПОРТНЫХ
бюджетовадминистративно-	вложениемкапиталоввценныебумаги

6.2. Wyrazy bez łącznika

Инв	индивидуализирующие
инвентаризировать	Индосамент
инвентаризируются	ИНДОССАМЕНТНЫЙ
инвентаризуются	ИНДОССО
инвойс	ИНКОТЕРМС
индентификации	

6.3. Przykładowe kategorie jednostek nierozpoznawalnych przez analizator

a) Błędy ortograficzne:

финасовых 8575 9 RLE aa -?? ФИНАСОВЫЙ йуйхйч -1 0

b) Skróty:

Сост 7916 4 RLE Aa NAM? -?? СОСТЫЙ йш -1 0

c) Wyrażenia złożone:

начала-середины 3740032 15 RLE aa -?? НАЧАЛО-СЕРЕДИНА гбгжгй -1 0
Аналитика-Пресс 7493 15 RLE Aa NAM? SENT1 -?? АНАЛИТИКА-ПРЕСС ааag -1 0

Jak widzimy, analizator podaje (proponuje) kody gramatyczne. Np. dla lematu НАЧАЛО-СЕРЕДИНА *Program* podaje kod *гбгжгй*, który odsyła nas do informacji, że dany wyraz zbudowany jest z kilku rzeczowników, czyli jest złożony:

Гб	С	жр,но	ед,рд	пальмы
Гж	С	жр,но	мн,им	пальмы
Гй	С	жр,но	мн,вн	пальмы

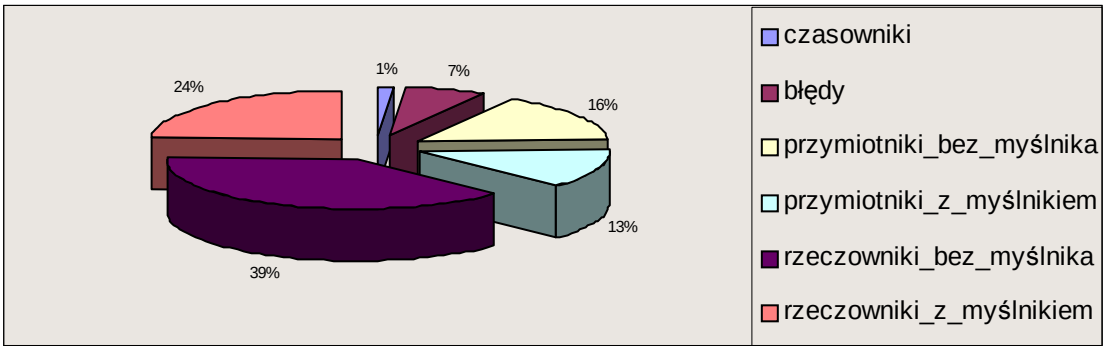
d) Przeniesienie wiersza, ręczny podział wiersza:
ПЕРСОНАЛИ-СТИЧЕСКАЯ 3741242 19 RLE AA -?? ПЕРСОНАЛИ-СТИЧЕСКИЙ йж -1 0

6.4. Kategorie części mowy – wyniki

АГЕНТ-ТРАНСФЕР 26823 14 RLE AA SENT1 -?? АГЕНТ-ТРАНСФЕР ааag -1 0
ведущуюся 100266 9 RLE aa -?? ВЕДУЩИЙСЯ йй -1 0
БЕЗНОМИНАЛЬНЫЕ 94287 14 RLE AA -?? БЕЗНОМИНАЛЬНЫЙ йтРь -1 0
Рудановского 475491 12 RLE Aa NAM? FAM2 -?? РУДАНОВСКИЙ йбйгйн -1 0
МСОК 1136425 4 RLE AA -?? МСОКИЙ йш -1 0
(skrót analizator traktuje jako jednostkę adjektywną)

70 procent (111 wyrazów) analizowanych wyrazów należy do kategorii przymiotników. Przymiotniki z łącznikiem zajmują 65 procent. 20 procent rzeczowników także należy do kategorii wyrazów zawierających łącznik. Słownictwo badanego Słownika wskazuje na składanie się istniejących już słów w wyrażenia podwójne, np. акциями-барометрами jako na tendencję rozwoju terminologii w dziedzinie księgowości. Końcowe wyniki przestawiają się następująco:

czasowniki	24
błędy	118
przymiotniki bez łącznika	282
przymiotniki z łącznikiem	217
rzeczowniki bez łącznika	663
rzeczowniki z łącznikiem	421



7. Podsumowanie

Konkretne wyniki kwantytatywne naszych analiz przedstawiliśmy we wcześniejszych partiach niniejszego artykułu. W tym miejscu skoncentrujemy się na sformułowaniu uwag natury ogólnej, dotyczących naszej pracy. Automatyczna analiza morfologiczna zajmuje niewątpliwie istotne miejsce we współczesnej praktyce językoznawczej. Analizator morfologiczny jako narzędzie lingwistyczne oferuje również inne możliwości oprócz np. sprawdzenia poprawności wyrazów. Są to m.in. analizy dotyczące własności gramatycznych (oznaczonych odpowiednimi kodami) analizowanych jednostek. W dalszej kolejności możliwe są do przeprowadzenia operacje na zbiorach kodów gramatycznych (w zależności od konkretnego celu pracy z tekstem). Opierając się na wynikach uzyskanych w niniejszym artykule, trzeba przyznać, że metoda analizy morfologicznej pozwala w znacznym stopniu uzyskać wiele (tj. na skalę masową) informacji językoznawczych. Warto jednak podkreślić, że uzyskane wyniki takiej analizy, np. dane frekwencyjne oraz zbiór wyrazów nierozpoznanych, nie mogą bezpośrednio prowadzić do formułowania wniosków dotyczących faktów językowych. Niezbędna w tym miejscu jest detaliczna analiza każdego pojedynczego wyniku analizy morfologicznej. Praca przeprowadzana z automatycznym analizatorem morfologicznym zawsze wymaga dodatkowej obserwacji lingwistycznej.

Bibliografia

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

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

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.

Linde-Usiekniewicz, J. (red.). 2002. *Wielki słownik angielsko-polski*. Warszawa: PWN.

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

Rabiega-Wiśniewska, J., Rudolf, M. 2002. AMOR – program automatycznej analizy fleksyjnej tekstu polskiego. *Biuletyn Polskiego Towarzystwa Językoznawczego* LVIII, 175 – 185.

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

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

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

Szafran, K. 1997. Automatyczne hasłowanie tekstu polskiego. *Polonica* XVIII, 51 – 64.

Wawrzyńczyk, J., Bolszowa, J., Radolińska, W., 1977. *Słowotwórstwo i fleksja współczesnego języka rosyjskiego*. Łódź: Wydawnictwo UŁ.

Wawrzyńczyk, J. (red.). 2004. *Wielki słownik rosyjsko-polski z Kluczem polsko-rosyjskim*. Warszawa: PWN.

Wawrzyńczyk, J., Małek, E. 2004. *Z materiałów do Słownika bibliograficznego języka rosyjskiego. Terminologia lingwistyczna. Wybrane terminy wiedzy o kulturze i literaturze. Neologizmy, hapaks legomena*. Warszawa: Semiosis lexicographica.

Wierzchoń, P. 2005. Automatyczne metody ekscerpcji neologizmów, czyli słowotwórstwo faktograficzne. *Scripta Neofilologica Posnaniensia* VII, 221 – 239.

Wojnowski, J. (red.) 2003. *Wielka encyklopedia PWN*. Warszawa: PWN.

Woliński, M., Przepiórowski, M. 2001. *Projekt anotacji morfosyntaktycznej korpusu języka polskiego*. Raport Nr 938, Warszawa. http://www.ipipan.waw.pl/~wolinski/publ/ipi938.pdf

Vetulani, Z. 2004. *Komunikacja człowieka z maszyną*. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Азрилиян, А.Н. 1999. *Большой бухгалтерский словарь*. Москва: Институт новой экономики.

Зализняк, А.А. 1987. *Грамматический словарь русского языка.* Москва: Русский язык.

APPENDIX

1. При лемматизации для каждого слова входного текста выдается множество морфологических интерпретаций следующего вида:

- лемма (всегда пишется большими буквами);
- морфологическая часть речи;
- набор общих граммем (которые относятся ко всем словоформам парадигмы слова).
- множество наборов граммем.

2. Gramem (граммема) to elementarny odsyłacz morfologiczny: odnosi słowoformę do danej klasy morfologicznej, np. do słowoformy *стол* z lematem СТОЛ będzie się odnosił następujący zbiór gramemów: "**мр, ед, им, но",** "**мр, ед, вн, но"**. W taki oto sposób analiza morfologiczna podaje dwa warianty analizy słowoformy *стол* z lematem СТОЛ wewnątrz jednej interpretacji morfologicznej: w Bierniku (**вн**) oraz w Mianowniku (**им**).

Przykłady gramemów: **мр, жр, ср** – мужской, женский, средний род; **од, но** – одушевленность, неодушевленность; **ед, мн** – единственное, множественное число; **им, рд, дт, вн, тв, пр** – падежи: именительный, родительный, дательный, винительный, творительный, предложный; **св, нс** – совершенный, несовершенный вид; **пе, нп** – переходный, непереходный глагол; **дст, стр** – действительный, страдательный залог; **нст, прш, буд** – настоящее, прошедшее, будущее время; **пвл** – повелительная форма глагола; **1л, 2л, 3л** – первое, второе, третье лицо; **0** – неизменяемое. **кр –** краткость (для прилагательных и причастий); **сравн –** сравнительная форма (для прилагательных; **имя, фам** – имя, фамилия; **лок, орг** – локативность, организация.; **кач –** качественное прилагательное; **вопр,относ –** вопросительность и относительность (для наречий); **дфст –** слово обычно не имеет множественного числа; **опч –** частая опечатка или ошибка; **жарг –** жаргонизм.