Magdalena Derwojedowa
Witold Kieraś
Danuta Skowrońska
Robert Wołosz

Автоматический морфологический анализ древних текстов. Современные орудия и анализ текстов II половины XIX века

DEC-2012/07/b/hs2/00570

Орудия для обработки текстов на польском языке

- орудия для современного польского языка:
  - создавание и обыск корпусов текстов *Narodowy Korpus Języka Polskiego* (http://nkjp.pl)
  - морфологические анализаторы (*Morfeusz*, http://sgjp.pl/morfeusz/; *Polimorf*, http://zil.ipipan.waw.pl/PoliMorf; *Słownik fleksyjny języka polskiego* na CD, POLENG)
  - семантически-лексические ресурсы (польский ворднет, *Słowosieć,* http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/slowosiecc, польский фраменет *RAMKI*, http://www.ramki.uw.edu.pl)

Орудия для обработки текстов на польском языке

- орудия для современного польского языка
  - синтаксические анализаторы (*Świgra*, http://nlp.ipipan.waw.pl/~wolinski/swigra/)
  - валентные словари (*Walenty*, http://clip.ipipan.waw.pl/Walenty)
  - составы орудий для работы с текстом (*PSI toolkit* http://psi-toolkit.wmi.amu.edu.pl/; орудия группы G4.19, http://nlp.pwr.wroc.pl/pl/narzedzia-i-zasoby/narzedzia-przetwarzania-morfosyntaktycznego)

Орудия для обработки текстов на польском языке

- орудия для анализа польского языка ранейших времен
  - *Korpus tekstów staropolskich* (http://www.ijp-pan.krakow.pl/publikacje-elektroniczne/korpus-tekstow-staropolskich)
  - *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do roku 1772)*
  - *Słownik polszczyzny XVI w.* (http://www.spxvi.edu.pl/).
  - *Słownik polszczyzny XVII i I poł. XVIII wieku* (http://xvii-wiek.ijp-pan.krakow.pl/pan_klient/).

Орудия для обработки текстов на польском языке

- создавание ресурсов, особенно корпусов, требует создания орудий для их анализа (извлечения данных).

- первым шагом в случае языковых данных является именно автоматический морфологический анализ.

Проект

Автоматический флективный анализ польских текстов с 1830 года до 1918 года с учетом изменений по изменениям слов и правописанию

2013–2016, DEC-2012/07/b/hs2/00570

http://www.f19.uw.edu.pl/

Создание флективного анализатора для польского языка второй половины новопольского срока имеет решительное значение для дальнейшего развития диахронических компьютерных исследований польского языка.

Проект

- Цели проекта:
  - разработка лексикографической концепции описания изменений по флексии и правописанию в (электронным) граматическим словаре.
  - разработка небольшого корпуса текстов с 1830 года до 1918 года.
  - разработка парадигматических образцов, неучтеных в *Грамматическом словаре польского языка*, но существующих в изменениях в исследоваемый срок, в формате, который сделает возможным автоматический морфологический анализ.
  - расширение словаря морфологического анализатора на заглавные слова, присутствующие в корпусе, но не опознаваемые анализатором.

Проект

- Результаты:
  - создание систематического, формализованного описания польской флексии II половины новопольского срока.
  - морфологический анализатор.
  - балансированный, маркированный и проверенный языковедами корпус текстов II половины новопольского срока.
  - систематика флективных категорий и их значений.
  - схема описания эволюционных явлений в электронном словаре.

Эксперимент

- предположение: анализатор с широким лексическим основанием может правильно анализировать значительный процент единиц как минимум в текстах от II половины XIX века.

- исключение «устарелых» или «старых», «уже не используемых» слов из лексикона орудий.

- создавание нового анализатора нецелесообразно, рекомендуемым решением является модификация существующего богатого анализатора.

Эксперимент

Słownik języka polskiego pod red. W. Doroszewskiego (Словарь польского языка; дальше SJPDor) в качестве основания списков заглавных слов для анализаторов Morfeusz (Морфеуш) и PoMor (ПоМор).

- SJPDor содержит лексику даже от II половины XVIII века.
- сверх 15 тыс. единиц были в SJPDor определены устарелыми и включены в Morfeusza-SGJP.

Эксперимент

- материал для исследований:
  - без малого 260 тыс. сегментов из романа «Кукла» Б. Пруса.
  - «Пан Тадеуш» А. Мицкевича

- итоги:
  - Морфеуш- SGJP не опознал без малого 4000 (без малого 1100 из них это уникальные слова) единиц.
  - использование анализатора ПоМор с возможностью учета ранеейшей лексики, уменшило состав неопознанных единиц в 12%.

Неопознанные единицы

- формы собственных имен
  - *Belcia*
  - *Obermann*
  - *Pantarkiewiczówna*
- иноязычные цитаты
  - *d'habits*
  - *gegangen*
  - *exposition*

Неопознанные единицы

- иная флексия
  - *ramiony* → *ramionami* (ramię:subst.instr.pl; плечи)
  - *fakta* → *fakty* (fakt:subst.nom.pl; факт)
  - *interesa* → *interesy* (interes:subst.nom.pl; дела)

Неопознанные единицы

- иная флексия
  - *nagiemi* → *nagimi* (nagi:adj.instr.pl; голый)
  - *zielonem* → *zielonym* (zielony:adj.instr.sg; зеленый)
  - *poplamionem* → *poplamionym* (poplamiony:adj.instr.sg; запятнанный)

Неопознанные единицы

- другая орфография
  - *imitacya →imitacja* (имитация)
  - *subjekt → subiekt* (продавец)
  - *wogóle → w ogóle* (вообще)
  - *zpewnością → z pewnością* (наверно)

Литературные стилизации в современных текстах

Melba z bananem –
Podejściem **wyszukanem**;
Sto gram z grzybkiem –
Szybkiem.

Ogden Nash, Candy is dandy; but liquor is quicker
(перевод S. Barańczak)

Расширение современного анализатора о прежние единицы

- Собрание корпуса в размерах 1 милиона сегментов который будет включать образцы длиной около 1000 сегментов, (для сожранения стилистического различия, образцы будут подбираться из сочинений разного типа (художественная проза, научно-технические публикации, драма, публицистика, газетные известия).

- Вступительный анализ (с помощью немодифицированного современного анализатора)
  и составление списка неопознанных единиц которые являются возможными древними формами.

- Обогащение состава форм заглавными словами «Варшавского Словаря», неотмечаемыми в современных словарях польского языка.

Расширение современного анализатора о прежние единицы

- Отнесение к единицам образцов изменений а при их отсутсвии – создание их.
- Всесторонний анализ данных, по итогам которого будут идентифицироваться а затем описываться образцы изменений, которые не существуют в современное время, если они будут отличаться от современного состояния каким-либо флективным параметром.
- Отмечание изменеий, касающихся характеристик лексем, напр. изменения парадигм изменений.
- Анализ ошибок и проверка качества новых принципов а также правильности всех отнесенных морфологических описаний.

Спасибо за внимание!

http://www.f19.uw.edu.pl/

Библиография

- Acedański S., *A Morphosyntactic Brill Tagger for Inflectional Languages*, [w:] *Advances in Natural Language Processing. 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16–18, 2010*, red.
- H. Loftsson, E. Rögnvaldsson i S. Helgadottir, s. 3–14, Springer 2010.
- Klemensiewicz Z., *Historia języka polskiego*, Warszawa 2002.
- Kwapień E., *Kształtowanie się zasobu leksykalnego polszczyzny XIX wieku*, Warszawa 2010.
- Maziarz M., Piasecki M. i Szpakowicz S., *Approaching plWordNet 2.0*, [w:] *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan 2012.
- Rudolf M., *Metody automatycznej analizy korpusu tekstów polskich: pozyskiwanie, wzbogacanie i przetwarzanie informacji lingwistycznych*, Wydział Polonistyki, Uniwersytet Warszawski 2004.
- Saloni Z., Woliński M., Wołosz R., Gruszczyński W. i Skowrońska D., *Słownik gramatyczny języka polskiego*, II wyd., Warszawa 2012, CD.
- Woliński M., *Komputerowa weryfikacja gramatyki Swidzińskiego*, Rozprawa doktorska, Instytut Podstaw Informatyki, Polska Akademia Nauk, Warsaw 2004.
- Woliński M., *Morfeusz — a Practical Tool for the Morphological Analysis of Polish*, [w:] *Intelligent Information Processing and Web Mining*, red. M.A.
- Kłopotek, S.T. Wierzchoń i K. Trojanowski, Advances in Soft Computing, s. 503–512, Springer-Verlag, Berlin 2006.
- Wołosz R., *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*, Warszawa 2005.