IV Международный конгресс
исследователей русского языка

РУССКИЙ ЯЗЫК:
исторические судьбы
и современность

ТРУДЫ
и МАТЕРИАЛЫ

Москва, МГУ
20–23 марта 2010 г.

Русский язык: исторические судьбы и современность

IV Международный конгресс исследователей русского языка

Москва, МГУ имени М. В. Ломоносова, филологический факультет

20-23 марта 2010 года

Труды и материалы

Lomonosov Moscow State University (MSU)
Faculty of Philology

Russian Language: Its Historical Destiny and Present State

The Fourth International Congress of Russian Language Researchers

Moscow, Lomonosov Moscow State University,
FACULTY OF PHILOLOGY

March 20-23, 2010

Proceedings and materials

Collected by

Marina L. Remneva, Anatoliy A. Polikarpov

Moscow University Press

2010

Московский государственный университет имени М. В. Ломоносова
Филологический факультет

Русский язык: исторические судьбы и современность

IV Международный конгресс исследователей русского языка

Москва, МГУ имени М. В. Ломоносова,
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

20-23 марта 2010 года

Труды и материалы

Составители
М. Л. Ремнёва, А. А. Поликарпов

Издательство Московского университета
2010

пертизы каждого типа омонимии. Несмотря на большой исторический возраст, данный метод для русского языка в полной мере не описан в открытой литературе, некоторые принципиальные идеи и реализация представлены в [1]. В [3] даны сравнительные оценки различных модулей разрешения омонимии, построенных на основе статистических методов и метода, основанного на правилах. В целом, оценки для случая полного разрешения функциональной омонимии достаточно близки 97,26% и 96,87%. Можно предположить, что неразрешенные примерно 3–5% относятся к синтаксически сложным случаям и многие авторы сходятся во мнении, что наиболее эффективным является использование гибридных технологий разрешения омонимии. Однако следует отметить, что полученные оценки даны для классификации типов омонимии, принятой в Национальном корпусе русского языка. Эта классификация в ряде конкретных случаев функциональной омонимии расходится с другими классификациями. Предварительно можно отметить, что развитие контекстного метода способствует более четкому выделению основных проблем, связанных, прежде всего, с описанием явления функциональной омонимии в существующих лексикографических источниках; выделением синтаксически сложных случаев разрешения омонимии.

Разработка метода контекстного разрешения функциональной омонимии [2] требует решения следующих задач:

1) лексикографические задачи:
- уточнение набора грамматических характеристик функциональных омонимов;
- построение полной классификации типов функциональных омонимов.

2) вычислительные задачи:
- выделение минимального множества разрешающих контекстов для каждого функционального типа. Формализация контекстных условий.
- для каждого функционального типа построение управляющей структуры обобщенного правила, обеспечивающего максимальную точность распознавания.

2. Основные результаты и проблемы метода контекстных правил

На основе сравнительных сопоставлений различных лексикографических источников, включая словари и корпуса русского языка, задача построения достаточно полного списка грамматических омонимов с уточненными грамматическими характеристиками близка к завершению. Связанная с этой задачей задача классификации типов функциональных омонимов также практически решена. В настоящее время выявлено около 220 классов грамматических омонимов (по оценкам Т. Ю. Кобзаревой в [1] — 57 классов) и построены списки представителей этих классов.

Для каждого типа функциональной омонимии разрабатывается обобщенное правило разрешения омонимии данного типа. Обобщенное правило представляет собой упорядоченную совокупность правил, записанных на специальном формальном языке. Каждое правило внутри совокупности фиксирует некоторый разрешающий контекст, порядок применения правил внутри функционального типа базируется на оценке частотности контекстов. Развитие метода связано с учетом контекстов сложной синтаксической природы, в частности, с анализом однородных групп. Выделение однородной группы позволяет искать разрешающий элемент за границами однородной группы; тем самым, реально увеличивается численный интервал разрешающего контекста. Такого рода правила анализа омонимов в составе однородной группы были включены в состав обобщенных правил различных функциональных типов.

Метод разрешения функциональной омонимии на основе контекстных правил по сути своей базируется на синтаксических моделях. Это обстоятельство определяет и ограничения метода. Приписывание омониму той или иной характеристики части речи осуществляется на основе анализа наличия либо отсутствия в контексте определенной длины слов тех или иных классов. Явления эллипсиса и субстантивации в тексте также представляют сложную проблему метода.

Одним из возможных подходов к разрешению омонимии в сложных (коротких и эллиптических) контекстах является логико-семантический подход к описанию предложений. Нами исследованы логико-семантические интерпретации омонимичных конструкций, относящихся к области неясных и спорных синтаксических явлений, а именно конструкций, находящиеся в тесной смысловой и синтаксической зависимости от окружающего контекста на примере функционального поведения омонимов типа N* / A*. Построена классификация сложных контекстов и предложены логико-семантические интерпретации для правил разрешения омонимии.

Анализ сложных контекстов позволил выделить определенные типы омонимичных контекстов, разрешение которых требует полного синтаксического анализа. Кроме сложных случаев постсинтаксического разрешения можно выделить контексты, не разрешаемые синтаксическими методами, т. е. сохраняющие многозначность.

Литература

1. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды междунар. конференции Диалог'2002. М., 2002. С. 258–268.
2. *Невзорова О. А., Зинькина Ю. В, Пяткин Н. В.* Метод контекстного разрешения функциональной омонимии: анализ применимости // Труды межд. конф. Диалог'2006. М., 2006. С. 399–402.
3. *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика-2005. http://company.yandex.ru/grant/list.xml.

Дискриминантный анализ стилей текстовых произведений

В. В. Поддубный, А. С. Кравцова

Томский государственный университет

pvv@inet.tsu.ru, askravtsova@gmail.com

Тексты, признаки стиля, частоты признаков, ранги частот, дискриминантный анализ, русская проза

Summary. The using of discriminant statistical analysis to decision of the problem of the comparison of the styles of the text products on the base of features frequencies is considered. It is offered the procedure of normalization of frequencies by way of the transition from relative frequencies to its ranks with the following nonlinear transformation of ones into gaussian values. On example of the analysis of texts of russian novels of 19–th century the discrimination of the author's styles on the frequencies of the using of 55 syntactic words is organized.

1. Рассматривается применение дискриминантного статистического анализа к решению проблемы сравнения стилей текстовых произведений на основе частотных признаков. Предлагается процедура нормализации признаков путем перехода от исходных признаков к их рангам с последующим нелинейным преобразованием в нормально распределенные величины. На примере анализа текстов русской художественной прозы XIX века проведена дискриминация авторских стилей текстов по частотам употребления 55 служебных слов.

2. Дискриминантный анализ [2] является одним из мощных инструментов математической статистики, позволяющий исследовать статистические различия классов объектов, относительно однородных внутри каждого класса. Применительно к текстовым произведениям дискриминантный анализ позволяет исследовать степень различия текстовых произведений по авторству, жанру и прочим группирующим признакам при различных наборах признаков стилей текстов (частот употребления служебных слов, наиболее употребительных слов, биграмм и т. п.).

3. При фиксированном (выбранном) наборе признаков стилей текстов каждый текст может быть представлен точкой в многомерном пространстве частот признаков или некоторых (в общем случае нелинейных) функций от них. Произведения одного автора (или одного жанра) группируются в относительно компактные сгустки точек (классы),

в общем случае достаточно сильно перекрывающиеся для разных авторов (или разных жанров). При сильном перекрытии классов задача различения классов (стилей текстов) в многомерном пространстве признаков стилей является достаточно трудной. Дискриминантный анализ позволяет максимально разнести классы друг относительно друга.

4. Пусть $p = \{p_{ij}\}$ — $n \times m$-матрица относительных частот появления j-го признака в i-м тексте ($i = 1,n$, $j = 1,m$, где n — число текстов, m — число признаков). Проранжировав в порядке возрастания величины $\{p_{ij}\}$ по всем текстам (от 1-го до n-го) для каждого j-го признака, получим матрицу рангов $r = \{r_{ij}\}$ признаков (мест признаков среди текстов). При этом рангам совпадающих частот, образующих так называемые связки, припишем средний по связке ранг. Разделив ранги на $n + 1$, приведем их к интервалу $[1 / (n + 1), n / (n + 1)]$. В результате получим матрицу относительных рангов $\{r_{ij} / (n + 1)\}$. Их эмпирическое распределение вероятностей равномерно в единичном интервале. Сопоставим каждому относительному рангу квантиль стандартного нормального распределения уровня этого относительного ранга. В результате такого нелинейного преобразования получим матрицу нормально распределенных величин (нормализованных относительных рангов — НОР) $x = \{x_{ij}\}$ с нулевыми средними и единичными дисперсиями для каждого j-го столбца, причем столбцы будут коррелированны (в общем случае) между собой.

5. Пусть имеется g классов. Вычислив положение центров классов в признаковом пространстве НОР (средние значения координат точек каждого k-го класса), можно подвергнуть оси координат такому линейному преобразованию $y = xV$ (повороту и масштабированию осей), при котором расстояния между центрами классов по отношению к диаметрам классов в новом (дискриминантном) признаковом пространстве станут наибольшими. Дискриминантный анализ предписывает выбирать матрицу V коэффициентов этого преобразования так, чтобы максимизировать отношение $\lambda_l = (V'BV)_{ll} / (V'WV)_{ll}$, $l = 1, 2, \ldots, q$, $q = \min(m, q - 1)$. Здесь штрих — знак транспонирования, $B = T - W$, T — $m \times m$-матрица ковариаций векторов-столбцов матрицы x, W — $m \times m$-матрица внутригрупповых ковариаций векторов-столбцов матрицы x. Из этого критерия оптимизации следует [1]; [2], что m-векторы-столбцы $\{V_l\}$ матрицы V — собственные векторы, соответствующие матрицам B и W и удовлетворяющие уравнению $BV_l = \lambda_l WV_l$, а $\{\lambda_l > 0\}$ — q их первых (в порядке убывания) собственных значений, удовлетворяющих характеристическому уравнению $\det(B - \lambda W) = 0$. Столбцы матрицы $y = \{y_{il}\}$ называются дискриминантными функциями и образуют новое признаковое пространство размерности q (пространство новых факторов), в котором обеспечивается наилучшее разделение (дискриминация) классов. Столбцы матрицы коэффициентов $V = \{V_{jl}\}$ называются факторными нагрузками. При соответствующей нормировке они являются коэффициентами корреляции между каждым новым l-м (дискриминантным) признаком и каждым «старым» j-м признаком пространства нормализованных относительных рангов. Содержательная интерпретация дискриминантных функций (новых, дискриминантных признаков) определяется наборами старых признаков, в наибольшей степени коррелирующими с новыми признаками. Статистическая значимость оставляемой в новом признаковом пространстве l-й дискриминантной функции при гауссовом распределении нормализованных относительных рангов рассчитывается по χ^2-распределению с $\nu_l = (m - l)(g - l - 1)$ степенями свободы, которому при верной нулевой гипотезе подчиняется величина, пропорциональная логарифму Λ-статистики Уилкса [2].

6. В качестве примера приводятся результаты дискриминантного анализа 81 текста крупных произведений художественной прозы (романов, повестей) 12 русских писателей XIX века с использованием в качестве признаков стилей 55 служебных слов. На рис. 1 видно отличное разделение текстов по писателям в пространстве первых двух дискриминантных функций.

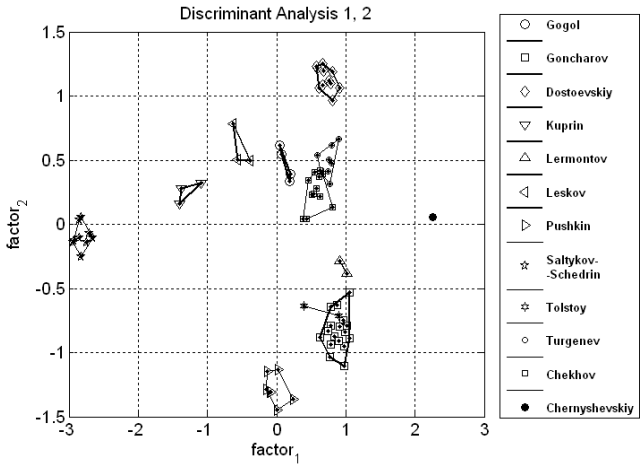

Рис. 1. Дискриминантный анализ текстов русских писателей XIX века по 55 служебным словам.

Литература

1. *Кендалл М. Дж., Стьюарт А.* Многомерный статистический анализ и временные ряды / Пер. с англ. М., 1976.

2. *Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р. и др.* Факторный, дискриминантный и кластерный анализ / Под ред. И. С. Енюкова; пер. с англ. М., 1989.

Математическое моделирование жизненного цикла языкового знака

В. В. Поддубный

Томский государственный университет

pvv@inet.tsu.ru

А. А. Поликарпов

Московский государственный университет имени М. В. Ломоносова

anatolp@philol.msu.ru

Языковой знак, жизненный цикл, полисемия, математическая модель, диссипативный процесс

Summary. The dissipative nonstationary dynamic mathematical model of the life cycle of the language sign is offered. This cycle is based on the interaction of the processes of the sign polysemy growing and of the losing of the earlier gained sign meanings. It is shown that this model satisfies to the variational principle of the least action. The model is presented in continuous and discrete variants. The stochastic expansion of the discrete variant of the model is built. The numerical modeling of the process of the language sign polysemy is made.

1. Жизненный цикл языкового знака от момента его зарождения до момента выхода из употребления связан с взаимодействием двух процессов его развития: процесса роста полисемии знака, приобретения знаком новых, как правило,