КОРПУСНОЕ ИССЛЕДОВАНИЕ СОЧЕТАЕМОСТНЫХ ПРЕДПОЧТЕНИЙ ЧАСТОТНЫХ ЛЕКСЕМ РУССКОГО ЯЗЫКА

CORPUS ANALYSIS OF SELECTIONAL PREFERENCES OF FREQUENT WORDS IN RUSSIAN

Митрофанова О.А. *(alkonost-om@yandex.ru),* ***Белик В.В.*** *(ogibbion14@pisem.net),*
Кадина В.В. *(veraiii@yandex.ru), Санкт-Петербургский государственный университет*

В докладе анализируются результаты исследования дистрибутивных свойств частотной лексики русского языка. Установлено решающее правило для выявления устойчивых сочетаний лексем с учётом коэффициента взаимной информации MI. Сочетаемостные предпочтения лексем определены в терминах морфологических классов и лексико-семантических признаков их синтагматических соседей.

1. Цели и задачи исследования

Информация о сочетаемостных предпочтениях слов, извлекаемая из корпусов текстов, играет важную роль при выполнении многих задач компьютерной лингвистики, среди которых автоматическая классификация лексики [Pekar, Staab 2003], разрешение неоднозначности [Resnik 1997], уточнение семантико-синтаксических моделей сочетаемости лексем в словарных базах данных [Азарова и др. 2005; Иорданская, Мельчук 2007; Ďurčo 2007], контрастивные исследования [Agirre et al. 2003] и пр.[1] Словари сочетаемости, построенные в результате компьютерной обработки больших корпусов текстов, представляют собой богатейший лингвистический ресурс [Гельбух и др. 2004]. В распоряжении лингвистов уже есть современные инструменты, предназначенные для исследования синтагматических свойств лексики и подключаемые непосредственно к корпусам. Например, существуют ресурсы (Sketch Engine[2]; Collocation Database, etc.[3]), которые позволяют не только ранжировать сочетания в соответствии с мерой их устойчивости, но также определять частеречную принадлежность, синтаксические и в некоторых случаях лексико-семантические признаки входящих в них слов [Lin 1999; Pala 2006]. Однако количественные критерии для выявления устойчивых сочетаний до сих пор недостаточно изучены.

В естественном языке существуют особые механизмы, которые регулируют комбинаторику лексических единиц текста на формальном и содержательном уровнях. Данные механизмы необходимо учитывать при моделировании понимания текста, в связи с этим функционирование сочетаний слов, тяготеющих к совместному употреблению, является объектом пристального внимания учёных в аспекте их статистической устойчивости, формальной и семантической связанности [Борисова 1995; Иорданская, Мельчук 2007; Ягунова 2006]. По всей видимости, анализ сочетаний слов с этих позиций даёт возможность детально исследовать их сочетаемостные предпочтения, модели взаимодействия их лексических значений, а также получить данные о контекстных маркерах значений.

Итак, цель обсуждаемого проекта – изучение дистрибутивных свойств частотной лексики в корпусах текстов русского языка, требующее решения ряда задач, среди которых:

сбор и интерпретация данных о сочетаемости лексических единиц в корпусах текстов с учётом различных параметров (взаимное расположение элементов контекстов – правосторонние и левосторонние синтагматические соседи исследуемых лексем; веса элементов контекстов в зависимости от их позиции по отношению к исследуемым лексемам; ширина контекстного окна и пр.);

получение количественных оценок силы связей лексических единиц в устойчивых сочетаниях; выявление и формулировка их сочетаемостных предпочтений с учётом морфологических классов и лексико-семантических признаков синтагматических соседей в контекстах.

---

[1] См. также материалы конференции CONTEXT: http://context-07.ruc.dk/CONTEXT07MainPage.html
[2] Sketch Engine: http://www.sketchengine.co.uk/; http://www.fi.muni.cz/~thomas/corpora/searches/index.htm
[3] Collocation Database, etc.: http://www.cs.ualberta.ca/~lindek/demos.htm

2. Методика определения сочетаемостных предпочтений лексем

Сочетаемостные предпочтения лексемы *X* можно выявить, определив *{a, b, c…}* – множество её потенциальных синтагматических соседей в контекстах и упорядочив их с точки зрения различных признаков (например, принадлежность к ЛСГ – глаголы движения, интеллектуальной деятельности и пр., существительные – названия природных явлений, транспортных средств и пр., морфологический класс – глаголы, наречия, прилагательные, местоимения и пр., синтаксическая функция – актанты, сирконстанты, атрибуты, и пр.). Множество потенциальных синтагматических соседей лексемы *X* формируется в результате анализа выборочных совокупностей контекстов её употребления в корпусах текстов.

Количественный критерий предпочтительности синтагматических соседей *{a, b, c…}* для слова *X* может быть задан с учётом какого-либо коэффициента ассоциативной связи элементов в сочетаниях (например, в биграммах). В исследованиях применяются различные меры – *MI*, *T*, *Log-Likelihood*, *Z*, X^2, и пр. [Church, Hanks 1990; Evert, Krenn 2001]. В нашем случае был использован коэффициент взаимной информации *MI*, определяемый для биграмм типа *yX / Xy*, где $y \in$ *{a, b, c…}* – коллокат (левый или правый сосед) базовой лексемы *X*.

Коэффициент *MI* позволяет оценивать силу ассоциативной связи внутри сочетания слов (между лексемой *X* и её соседом *y*) на основе соотношения частоты встречаемости биграммы *f(X,y)* и независимых употреблений коллокатов *f(X)* и *f(y)*, с учётом объема корпуса *N*:

$$MI = \log_2 \left\{ \frac{N \bullet f(X,y)}{f(X) \bullet f(y)} \right\}$$

Чем выше значение коэффициента *MI*, тем более предпочтителен тот или иной синтагматический сосед *y* для лексемы *X* (и тем вероятнее, что *y* является маркером какого-либо из значений, закреплённых за *X*).

По сравнению с показателем частотности независимого употребления соседей лексемы *X*, коэффициент *MI* позволяет различать коллокаты с широкими сочетаемостными возможностями (которые могут оказаться высокочастотными и, вместе с тем, несущественными для лексемы *X*) и коллокаты, тяготеющие к употреблению в сочетаниях с лексемой *X* (и поэтому значимым образом характеризующие её сочетаемостные предпочтения).

Известно, что при извлечении биграмм из корпуса текстов с учётом значения *MI* удаётся выявить наибольшее число сочетаний, зарегистрированных в лексикографических источниках; доля биграмм со знаками пунктуации в экспериментах с *MI* оказывается существенно ниже, чем при использовании других мер, в частности, *T* и *Log-Likelihood* [Khokhlova 2007].

При формулировке сочетаемостных предпочтений слов предлагается использовать в качестве эвристики аппарат теории оптимальности, которая помогает смоделировать конкуренцию правил и ограничений, задействованных в построении языковых выражений на уровнях от фонологического до семантического [Blutner et al. 2006]. В зависимости от степени важности, эти правила и ограничения получают ранг. Чем важнее правило или ограничение, чем выше его ранг, тем серьёзнее его нарушение и тем менее «правильным» будет порождаемое языковое выражение. Правила и ограничения, имеющие низкий ранг, могут быть нарушены без ущерба для допустимости итогового языкового выражения. Иными словами, использование теории оптимальности в лингвистическом моделировании позволяет перейти от идеальных языковых структур к оптимальным (приемлемым в той или иной мере). С этих позиций можно установить иерархию приоритетов, существующих при выборе для лексемы *X* её синтагматических соседей с тем или иным лексическим значением, представляющих ту или иную часть речи, выполняющих ту или иную синтаксическую функцию.

3. Лингвистический материал, источники данных, исследовательские инструменты

Исследование проводится на материале наиболее частотных лексем русского языка, среди которых глаголы *идти, видеть, говорить, знать, сказать, есть, хотеть* и пр.; существительные *человек, год, рука, век, жизнь, друг, глаз* и пр.; прилагательные *близкий, далёкий, долгий, молодой, поздний, соседний, старший* и пр. Информация о сочетаемостных свойствах данных слов в дальнейшем использовалась при анализе контекстов их употребления в различных значениях. В ходе экспериментов осуществляется обработка лингвистических данных, содержащихся в ряде корпусных ресурсов: электронная библиотека М. Мошкова; корпус текстов русского языка Бокрёнок, применяемый на кафедре математической лингвистики СПбГУ; выборки типовых контекстов из Словаря русского языка С.И. Ожегова в формате базы данных Starling. Извлечение сочетаний слов производится с помощью сервиса поиска биграмм в лингвистическом ресурсе АОТ, где в качестве корпуса используется текстовая база электронной библиотеки М. Мошкова [Аверин 2006].[4] Данный сервис позволяет получать списки

[4] Сервис поиска биграмм в лингвистическом ресурсе АОТ: http://aot.ru/demo/bigrams.html

биграмм с левосторонними / правосторонними коллокатами ключевого слова, упорядоченные по значению *MI*, по частоте биграммы или частотам коллокатов.

4. Формулировка решающего правила для выявления устойчивых сочетаний слов

Для содержательной обработки сочетаемостных данных необходимо сформулировать решающее правило, помогающее выявлять устойчивые сочетания, а главное, требуется определить соответствующее пороговое значение коэффициента взаимной информации *MI* в биграммах. Известно, что для сочетаний языковых единиц разных типов этот показатель должен подбираться индивидуально [Азарова и др. 2005]. Так, например, для английского языка с фиксированным порядком слов установлено пороговое значение *MI* = 3 [Church, Hanks 1990]. Можно допустить, что для русского языка эта величина будет несколько ниже, поскольку в русскоязычных текстах преобладает свободный порядок слов.

При определения порогового значения *MI* в качестве эвристики использовался метод минимального риска, или минимакса [Джонсон, Лион 1980: 433–435]. Суть метода заключается в том, что в процессе установления принадлежности какого-либо объекта к некоторому классу противопоставляются три типа решений: «попадание в цель» (правильное решение), «ложная тревога» (инородный объект ложно квалифицируется как входящий в класс) и «промах» (объект из класса не распознается как принадлежащий к классу). Правильные решения поощряются дополнительными очками или весами. Также производится взвешивание ошибок: менее серьёзные ошибки – «ложные тревоги» – получают меньший вес; более серьёзные ошибки – «промахи» – получают больший вес. В рассматриваемом случае трактовка устойчивого сочетания как неустойчивого следует считать «промахом», а обратную ситуацию – «ложной тревогой». Иллюстрацией «промаха» может служить игнорирование сочетаний со знаменательными словами, являющимися маркерами лексического значения базовой лексемы: например, *MI* (*говорить* + *язык*) = 0,777. «Ложные тревоги» чаще всего возникают в сочетаниях базовой лексемы и незнаменательных слов – местоимений, союзов, реже предлогов: например, *MI* (*говорить* + *я*) = 1,262.

При анализе биграмм было обнаружено, что оптимальное соотношение «попаданий в цель», «промахов» и «ложных тревог» достигается при *MI* = 1. В среднем, доля правильных решений составляет 87%, на десять «попаданий в цель» (вес «3») приходится один «промах» (вес «2») и две «ложных тревоги» (вес «1»). Изменение порогового значения приводит к снижению доли правильных решений и к увеличению доли ошибок. Таким образом, искомое решающее правило имеет следующий вид:

- если $MI \geq 1$, то сочетание слов считается устойчивым;
- если $MI < 1$, то сочетание слов оценивается как неустойчивое.

Расширенная версия данного решающего правила, учитывающая критерии для выявления связанных сочетаний различных типов (свободные / связанные, квазифраземы (коллокации) / фраземы, квазиидиомы / идиомы: согласно классификации, описанной в [Иорданская, Мельчук 2007]), приведена в работе [Митрофанова 2008].

Для верификации решающего правила было произведено сравнение результатов анализа биграмм, содержащих частотные лексемы русского языка, и информации об их синтагматических соседях, полученной в ходе ручной обработки представительных выборок из корпуса Бокрёнок [Митрофанова и др. 2006], а также типовых контекстов из Словаря русского языка С.И. Ожегова в формате базы данных Starling (СО) [Митрофанова, Крылов 2006]. Оказалось, что практически все синтагматические соседи, выявленные в контекстах из корпуса, встречаются в биграммах с $MI \geq 1$ (точнее, $MI \in [1, 3]$). Немногочисленным идиомам соответствуют биграммы с ещё более высоким значением *MI* (ср. *MI (речь + идти)* = 7,495; *MI (идти + вразрез)* = 9,466 и пр.)

Например, при интерпретации данных об употреблении существительного *человек* были обнаружены устойчивые сочетания, фигурирующие и в типовых контекстах из СО, и в контекстах из корпуса Бокрёнок, и в биграммах, при этом значение *MI* выше порогового:

MI (*молодой* + *человек*) = 6,339;
MI (*первобытный* + *человек*) = 5,645;
MI (*честный* +*человек*) = 5,212;
MI (*разумный* +*человек*) = 3,886;
MI (*хороший* + *человек*) = 2,536;
MI (*природа* + *человек*) = 2, 453;
MI (*честный* + *человек*) = 2,359;
MI (*жизнь* + *человек*) = 1,643;
MI (*отношение* + *человек*) = 1,112.

Также были рассмотрены другие сочетания существительного *человек* с левосторонними и правосторонними синтагматическими соседями, встретившиеся в контекстах из корпуса Бокрёнок и зарегистрированные в

биграммах. Учитывалось положение соседей в сочетаниях, а также их тип с точки зрения решающего правила.

Левосторонние синтагматические соседи:

«попадания в цель»: прилагательные *здравомыслящий, порядочный, молодой, взрослый, умный, добрый, здоровый, простой, хороший, русский, живой, счастливый, близкий* и пр. количественные слова *миллиард, миллион, тысяча* и пр.; существительные *природа, сознание, судьба, жизнь, душа, сердце, мир* и пр.;

«ложные тревоги»: *этот, между, когда* и пр.; «промахи»: *любить, образ, имя* и пр.

Правосторонние синтагматические соседи:

«попадания в цель»: прилагательные *умный, добрый* и пр.; глаголы *обладать, иметь, жить, погибнуть, создать, начинать, заниматься, работать, уметь, пользоваться, сидеть, стоять, ходить, называть, считать, являться* и пр.;

«ложные тревоги»: *вообще, с, среди, ибо, то* и пр.; «промахи»: *нужно, хороший, молодой* и пр.

Тем самым, подтверждается предположение о том, что при выборе порогового значения *MI* = 1 удаётся учесть подавляющее большинство устойчивых сочетаний, при этом доля ошибок невелика.

5. Эксперименты по выявлению сочетаемостных предпочтений лексем в биграммах

Исследовательская процедура иллюстрируется на примере обработки биграмм с глаголом *идти* и прилагательным *далёкий*. Ниже приводятся фрагменты списков биграмм для данных слов, примеры их коллокатов в биграммах с $MI \geq 1$, сгруппированные на основе общности их морфологических и, где возможно, лексико-семантических признаков (таблицы 1–4). Данная информация была использована при формулировке и ранжировании сочетаемостных предпочтений изучаемых лексем.

5.1. Сочетаемость глагола идти

Биграммы, включающие глагол *идти* и левый контекст:

MI (*неторопливо* + *идти*) = 4,668;
MI (*смело* + *идти*) = 4,467;
MI (*поезд* + *идти*) = 4,278;
MI (*тропа* + *идти*) = 4,254;
MI (*надо* + *идти*) = 3,154;
MI (*решить* + *идти*) = 2,088;
MI (*мочь* + *идти*) = 1,770; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	скорость, время	*торопливо, неторопливо, медленно, быстро, долго* и пр.
	направление	*кругом, следом, впереди, навстречу, далее, далеко, куда, куда-то, некуда* и пр.
	эмоциональна оценка	*уверенно, смело, упорно* и пр.
существительные	средство передвижения	*караван, поезд, пароход* и пр.
	путь	*дорога, тропа* и пр.
	природное явление	*дождь, снег* и пр.
	сложное действие/процесс	*бой, разговор, торговля* и пр.
глаголы (в т.ч. предикативы)	∞	*мочь, продолжать, отказываться, молча, разрешить, решить, собираться, надо, пора* и пр.

Таблица 1. Группы левосторонних коллокатов глагола идти

Биграммы, включающие глагол *идти* и правый контекст:

MI (*идти* + *пешком*) = 7,240;
MI (*идти* + *ожесточённый*) = 5,475;
MI (*идти* + *далёкий*) = 4,642;
MI (*идти* + *спать*) = 3,860;
MI (*идти* + *отдыхать*) = 3,670;
MI (*идти* + *разговор*) = 2,175;
MI (*идти* + *волна*) = 1,218; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	противопоставление	*напролом, наперекор* и пр.
	средство, способ	*пешком, босиком* и пр.
	направление	*рядом, впереди, следом, навстречу, домой, вдоль, параллельно, кругом, вперёд, прямо, напрямик, наверх, мимо, сюда, туда* и пр.
	положительная оценка	*нормально, гладко* и пр.
прилагательные (препозитивные определения в зависимых именных группах)	интенсивность	*ожесточённый, непрерывный* и пр.
	скорость	*медленный, быстрый* и пр.
	расстояние	*далёкий, близкий* и пр.
глаголы	∞	*завтракать, гулять, спать, отдыхать, идти* и пр.
существительные	природное явление	*дождь, пар, снег, волна* и пр.
	сложное действие/процесс	*подготовка, спор, бой, разговор* и пр.

Таблица 2. Группы правосторонних коллокатов глагола идти

5.2. Сочетаемость прилагательного далёкий

Биграммы, включающие прилагательное *далёкий* и левый контекст:
MI (*бесконечно* + *далёкий*) = 6,631;
MI (*донестись* + *далёкий*) = 3,823;
MI (*пробираться* + *далёкий*) = 3,804;
MI (*весьма* + *далёкий*) = 3,748;
MI (*немного* + *далёкий*) = 3,656;
MI (*вершина* + *далёкий*) = 2,279;
MI (*чужой* + *далёкий*) = 1,251; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
наречия	мера, степень	*бесконечно, страшно, весьма, немного, столь, настолько, очень, слишком, более, довольно* и пр.
прилагательные	∞	*невообразимый, далёкий, чужой, самый, такой, какой-нибудь* и пр.
глаголы	∞	*пробираться, донестись, послышаться, услышать* и пр.
существительные	путь	*путь*, *дорога* и пр.
	место	*страна, край, берег, вершина* и пр.

Таблица 3. Группы левосторонних коллокатов прилагательного далёкий

Биграммы, включающие прилагательное *далёкий* и правый контекст:
MI (*далёкий* + *прошлое*) = 6,592;
MI (*далёкий* + *предок*) = 6,181;
MI (*далёкий* + *звезда*) = 4,734;
MI (*далёкий* + *окраина*) = 4,223;
MI (*далёкий* + *галактика*) = 4,111;
MI (*далёкий* + *далёкий*) = 3,334;
MI (*далёкий* + *южный*) = 2,291; и пр.

Морфологические классы	**Лексико-семантические признаки**	**Примеры**
существительные	время	*прошлое, будущее, предок, потомок, детство, юность, древность*
	место	*родина, даль, край, окраина, страна, планета, звезда, галактика* и пр.
	сложное действие/процесс	*путешествие, плавание* и пр.
	природное явление, звук	*раскат, гром, эхо* и пр.
прилагательные	расстояние	*далёкий, прошлый* и пр.
	место, направление	*горный, северный, южный и пр.*

Таблица 4. Группы правосторонних коллокатов прилагательного далёкий

5.3. Формулировка сочетаемостных предпочтений для лексем идти *и* далёкий

На основе информации о коллокатах лексем *идти* и *далёкий* удалось выявить морфологические модели типа *POS + X* / *X + POS* и ранжировать их в соответствии с наибольшими показателями *MI* в соответствующих группах биграмм. Рассматривались и другие способы ранжирования моделей (среднее геометрическое, мода), однако они не были достаточно эффективными.

Сочетаемостные предпочтения глагола *идти*:
ранг 1. *X + Adv*
ранг 2. *X + Adj*
ранг 3. *Adv + X*
ранг 4. *Noun + X*
ранг 5. *X + Verb*
ранг 6. *X + Noun*
ранг 7. *Verb + X*

Сочетаемостные предпочтения прилагательного *далёкий:*
ранг1. *X + Noun*
ранг 2. *Adv + X*
ранг 3. *X + Adj*
ранг 4. *Verb + X*
ранг 5. *Adj + X*
ранг 6. *Noun + X*

В ряде случаев оказалось возможным также сформулировать сочетаемостные предпочтения лексем в терминах лексико-семантических признаков их коллокатов, например, в сочетаниях типа *Adj + N*:

ранг 1. *Adj (далёкий) + N (время)*
ранг 2. *Adj (далёкий) + N (природное явление, звук)*
ранг 3. *Adj (далёкий) + N (место)*
ранг 4. *Adj (далёкий) + N (сложное действие/процесс)*
ранг 5. *N (место)+ Adj (далёкий)*
ранг 6. *N (путь) + Adj (далёкий)*

6. Итоги исследования и направления дальнейшей работы

В ходе исследования подтверждена возможность описания сочетаемостных предпочтений лексем на основе статистико-комбинаторных данных, извлекаемых из корпусов текстов, установлено решающее правило для выявления устойчивых сочетаний частотных лексем русского языка с учётом коэффициента взаимной информации MI. Сочетаемостные предпочтения рассматриваемых лексем определены в терминах морфологических классов и лексико-семантических признаков их коллокатов. Данная информация была использована в прикладных разработках: при анализе предикатно-аргументных структур в экспериментальном корпусе контекстов для частотных глаголов русского языка (кафедра математической лингвистики СПбГУ); в процессе создания словарных статей и подбора иллюстративных примеров употребления частотных прилагательных в проекте «Современный толковый словарь живого русского языка» (лаборатория компьютерной лексикографии СПбГУ).

В дальнейшем планируется:

- перевести процедуру выявления сочетаемостных предпочтений слов в полуавтоматический режим;
- произвести статистическую оценку эффективности метода выявления сочетаемостных предпочтений;
- дать лингвистическую и статистическую интерпретацию ошибочных решений;
- исследовать зависимость степени устойчивости сочетаний слов от синтаксической организации текста;
- применить разработанные описания сочетаемостных предпочтений лексем в процессе обучения компьютерного инструмента для разрешения лексико-семантической неоднозначности;
- осуществить эксперименты по разрешению лексико-семантической неоднозначности слов в русскоязычных текстах с учётом сочетаемостной информации.

Список литературы

1. Аверин А.Н. Разработка сервиса поиска биграмм // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 5–15.

2. Азарова И.В., Синопальникова А.А., Смрж П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2005». М.: 2005. С. 11–17.

3. Борисова Е.Г. Коллокации. Что это такое и как их изучать. М., 1995.

4. Гельбух А.Ф., Сидоров Г.О., Эрнандес-Рубио Э., Чубукова М.В. Словари сочетаемости слов: какой метод составления лучше? // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2004». М.: 2004. URL: www.dialog-21.ru/Archive/2004/Gelbukh.pdf

5. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Т. 1. Методы обработки данных М.: 1980.

6. Иорданская Л.Н., Мельчук И.А. Смысл и сочетаемость в словаре. М.: 2007.

7. Митрофанова О.А., Кадина В.В., Савицкий В.С. Словарь и корпус как источники данных о синтагматических связях лексических единиц // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 271–281.

8. Митрофанова О.А., Крылов С.А. «Типовой» контекст: случайность или закономерность? // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 382–388.

9. Митрофанова О.А. О решающем правиле для определения устойчивости и связанности сочетаний слов // Четвёртая научно-практическая конференция «Прикладная лингвистика в науке и образовании». СПб.: 2008 [в печати].

10. Ягунова Е.В. Неоднословные целостности в словаре и корпусе // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 395–412.

11. Agirre E., Aldezabal I., Pociello E. A Pilot Study of English Selectional Preferences and Their Cross-Lingual Compatibility with Basque // Text, Speech and Dialogue: 6th International Conference TSD–2003. Lecture Notes in Artificial Intelligence. Vol. 2807. Springer-Verlag: 2003. P. 12–19.

12. Blutner R., de Hoop H., Hendriks P. Optimal Communication. CSLI Lecture Notes. Vol. 177. Stanford: 2006.

13. Church K.W., Hanks P. Word Association Norms, Mutual Information, and Lexicography // Computational Linguistics. Vol. 16. 1990. P. 22–29.

14. Ďurčo P. Collocations in Slovak (Based on the Slovak National Corpus) // Computer Treatment of Slavic and East European Languages: 4th International Seminar. Bratislava: 2007. P. 43–50.

15. Evert S., Krenn B. Methods for the Qualitative Evaluation of Lexical Association Measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse: 2001. P. 188–195.

16. Khokhlova M. Collocations in Russian: Analysis of Association Measures // Computer Treatment of Slavic and East European Languages: 4th International Seminar. Bratislava: 2007. P. 96–103.

17. Lin D. Automatic Identification of Non-compositional Phrases // Proceedings of ACL–99. University of Maryland: 1999. P. 317–324.

18. Pala K. Word Skteches and Semantic Roles // Труды Международной конференции «Корпусная лингвистика – 2006». СПб.: 2006. С. 307–317.

19. Pekar V., Staab S. Word Classification Based on Combined Measures of Distributional and Semantic Similarity // Proceedings of European Chapter of ACL–03, Research Notes Session. Budapest: 2003. P. 147–150.

20. Resnik P. Selectional Preference and Sense Disambiguation // Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington: 1997. P. 52–57.