

**Применение «поискового образа» документа
из цифровой библиотеки музейного сайта
для упрощения обнаружения в Интернете самого документа**

**Applying the Document Search Image from
a Museum Site Digital Library to Simplify
the Document Retrieval on the Internet**

**Застосування «пошукового образу» документа
із цифрової бібліотеки музейного сайту
для спрощення знаходження в Інтернеті власне документа**

П. П. Макагонов, С. Р. Еспиноса

Технологический университет Микстек, Уахуапан-де-Леон, Мексика

К. О. Сбойчаков

Государственная публичная научно-техническая библиотека России, Москва, Россия

Pavel Makagonov and Celia Reyes Espinosa

Mixtec Technological University, Huajuapán de León, Mexico

Konstantin Sboychakov

Russian National Public Library for Science and Technology, Moscow, Russia

П. П. Макагонов, С. Р. Еспіноса

Технологічний університет Мікстек, Уахуапан-де-Леон, Мексика

К. О. Сбойчаков

Державна публічна науково-технічна бібліотека Росії, Москва, Росія

Для южных регионов Мексики и, в частности, микстекского региона штата Оахака, актуально создание центров культуры, региональных музеев, любой деятельности, направленной на сохранение культурного наследия, на разработку материалов дидактического характера на начальных школ по истории и культуре региона. В частности, актуально создание веб-страниц в интернете, являющихся виртуальным аналогом культурных центров. В работе рассматривается новое приложение разработанной ранее методики построения проблемно-ориентированных словарей для электронной библиотеки музейного сайта. Документы из библиотеки музейного сайта представляются в виде поискового образа – совокупности слов из документа, представленных в проблемно-ориентированном словаре. В связи с отсутствием в регионе традиции составления рефератов статей поисковые образы могут быть использованы как представители этих документов, пригодные для замещения ими рефератов в полнотекстовых базах и в поисковых системах, когда доступ к самим документам ограничен. Соответствие поисковых образов документов, их заголовков и текстов самих документов (статей и разделов книг) позволяет вести поиск в Интернете текстовых документов на сайте более эффективно по сравнению с поиском по заголовкам статей.

Ключевые слова: предметно- ориентированный словарь, поисковый образ текстового документа, электронная библиотека, сайт музея.

Работа выполнена при поддержке проекта CONACYT 51777.

South Mexican regions, and the Mixtec region of Oaxaca state in particular, is currently characterized by the development of cultural centers, regional museums, various cultural heritage preservation activities or development of methodological educational materials on local history and culture for primary schools. The creation of web pages as virtual analogs of cultural centers is widespread. The paper describes a new application of a previously developed methodology of building problem-oriented dictionaries for a museum site digital library. Documents from the museum site library are represented as search images, a combination of words from the document that are present in the problem-oriented dictionary. As there is no tradition of writing paper abstracts in the region, these search images can be used as document representatives able to substitute abstracts in full text databases and search engines when there is a limited access to the documents themselves. Correspondence between document search images and their titles on one side and the document texts (articles and book sections) on the other side allows for more productive full text search on the web as compared with title search.

Keywords: subject-oriented dictionary, search image of a text document, digital library, museum site.

This work has been supported by CONACYT 51777 project.

Для південних регіонів Мексики, в т.ч. мікстекського регіону штату Оахака, актуальним є створення центрів культури, регіональних музеїв, будь-якої діяльності, що спрямована на збереження культурної спадщини, на розробку матеріалів дидактичного характеру для початкових шкіл з історії та культури регіону. Зокрема, актуальним є створення веб-сторінок в інтернеті, які є віртуальним аналогом культурних центрів. У роботі розглянуто новий додаток розробленої раніше методики побудови проблемно-орієнтованих словників для електронної бібліотеки музейного сайту. Документи з бібліотеки музейного сайту подано у вигляді пошукового образу – сукупності слів із документу, які містяться у проблемно-орієнтованому словнику. У зв'язку з відсутністю у регіоні традиції реферування статей, пошукові образи можуть бути використані як представники цих документів, які підходять для заміщення ними рефератів у повнотекстових базах і в пошукових системах, коли доступ до самих документів є обмеженим. Відповідність пошукових образів документів, їх заголовків та текстів власне документів (статей та розділів книжок) дозволяє проводити пошук в Інтернеті текстових документів на сайті більш ефективно у порівнянні з пошуком за заголовками статей.

Ключові слова: предметно-орієнтований словник, пошуковий образ текстового документа, електронна бібліотека, сайт музею.

Роботу виконано за підтримки проекту CONACYT 51777.

1. Введение. Актуальность создания музейного сайта и электронной библиотеки по проблемам культуры в регионе Микстека

В последнее время обнаружилась потребность у населения Микстеки (региона штата Оахака, Мексика) в материалах по проблемам культуры своего региона. Последнее объясняется интенсивным процессом обезлюдения региона: в то время, как в Микстеке проживает около 550 тысяч коренного индейского населения, в США на заработки перебралось около 440 тысяч. Как у оставшейся части, так и у эмигрантов в США возникла тяга к своей культуре. В селениях региона стали создаваться общественные музеи по местной инициативе и, иногда, при материальной поддержке эмигрировавших земляков, иногда при спонсорстве государственных структур, ответственных за сохранение культурного наследия индейцев.

Существует также проблема преподавания истории, географии и культуры родного края в начальной школе, поскольку в сельской местности большой процент неграмотного взрослого населения, часто не владеющего испанским языком, а учителя – люди из других, продвинутых регионов (в основном, из центральных и северных штатов) и они не владеют информацией по культурно – историческому наследию региона и ближайшего окружения поселения. Школьные учебники такой информации не содержат.

В этих условиях была разработана методика создания «успешных» (популярных) сайтов для тех региональных музеев, основная функция которых – играть роль культурных, образовательно-просветительских, научных (археологических) и туристических центров [1, 2]. На базе этой методики создан сайт www.mureh.org.mx регионального музея MureH города Уахаупан де Леон (Huajuapán de León). За 7 месяцев сайт вышел на 9 место в поисковой машине Google при задании на поиск по словам «museo regional» на испанском языке, но без каких-либо слов, указывающих регион. Эта позиция соответствует принятому в предложенной методике критерию успешности некоммерческих сайтов. Сайт ориентирован на три типа пользователей: специалистов по культуре Мезоамерики, на молодежь региона и на эмигрантов из региона (в основном, в США).

За семь месяцев с момента открытия сайта зарегистрировано 3113 посещений, из них 63% – вышли на сайт через поисковые машины, а 21% непосредственно задавая адрес сайта.

В таблице 1 дано процентное распределение посещений сайта за первые семь месяцев. Это распределение служит отправным пунктом доказательства правильности исходных представлений о модели пользователя сайта.

Таблица 1.

Регион Мексики	% посетителей	Страны	% посетителей
Мехико (федеральный округ)	26. 24	США	9. 71
Штат Оахака и город Уахуапан	31. 80	Испаноязычные страны	7. 32
Остальные мексиканские штаты	23. 32	Остальной мир	1. 61

По данным социальной сети «El Directorio Oaxaqueño» о количестве этнических микстеков как активных пользователей интернета в каждом штате США выяснилось, что процент посещений испаноязычного музейного сайта в каждом штате США и количество микстеков в сети коррелируют с коэффициентом 0.966. Вместе с распределением, данным в таблице 1 этот факт служит опорным пунктом доказательства правильности начальных представлений о модели пользователя сайта.

Созданный сайт оказался не только ВЭБ приложением для обслуживания пользователей, но и инструментом для обратной связи с ними. Мониторинг посещений сайта дал словарь запросов – список поисковых слов по запросам, приведшим на сайт MugeN – позволяющий определить «социальный заказ» основных пользователей и отследить изменение тематики запросов во времени. Первый вариант *«словаря запросов»*, состоящий из семисот значимых слов был использован в дальнейшей работе.

Созданный сайт оказался также полезным накопителем информации по микстекской культуре, заполняющим в значительной мере культурное информационное пространство учебного процесса в начальных школах региона. После создания сайта к работе стали подключаться добровольцы – собиратели материалов по истории и культуре Микстеки. Таковыми оказались: учителя начальных школ отдельных муниципальных образований, директора общественных музеев, исследователи культуры региона. Поэтому наряду с сайтом музея пришлось создавать портал, посвященный культуре Микстеки, www.cumix.org.mx. Отдельные разделы портала находятся на разных этапах разработки и включают в себя следующие сайты: Библиотека; Музыка, песни и танцы региона; Гастрономическая культура; Виртуальные курсы микстекского языка; Текущие события культурной жизни; Ремесла; «Знаменитые микстеки», а также Сайты шести музеев, двух археологических зон, двух участков наскальных росписей.

Каждый раздел требует разного программного обеспечения и разрабатывается учетом необходимости сокращения времени загрузки в условиях «медленной сети».

Все разделы содержат текстовую информацию по конкретным районам и селениям Микстеки. Однако только библиотека организована как отдельная текстовая база данных.

Важными для региона пользователи этой информации являются учителя и ученики начальной школы. Поскольку и те и другие не имеют регулярного доступа к Интернету, то пришлось создавать версии сайта на компакт-дисках и распространять их по школам.

Помимо этого вне сайта в интернете имеется большое число документов, содержащих обобщенные начальные сведения по микстекской истории (доиспанский период), этнографии, археологии, традициях и по современной культуре.

Эта информация была собрана в электронную библиотеку сайта с различной дисциплиной доступа к каждому документу в зависимости от желания автора материала и с учетом соблюдения его прав.

2. Поисковый образ текста в библиотеке сайта, полученный на основе проблемно-ориентированного словаря

В качестве поисковых средств в библиотеках, коллекциях текстов выступают названия книг, заголовки статей и разделов книг, индексы слов предметной области, рефераты и ключевые слова статей. Заголовки в силу ограничений их объема (такие ограничения часто определяются редакциями журналов и конференций) не содержат части важной информации о содержании основной части статьи или доклада. Рефераты составляются авторами иногда формально, часто повторяют постановку задачи и результаты, но не пути их достижения, раскрытые в основной части статьи. Иногда содержание реферата составляется с учетом специфики конференции, на которую представляется доклад. Такое же влияние оказывает профиль журнала или тема конференции на отбор авторами ключевых слов. Индекс полного собрания заголовков докладов Вольного экономического общества России[3], например, содержит на 50% слова, встречающиеся всего один раз и множество слов, встречающихся более 20 раз. В обоих случаях эти слова фактически не работают как инструмент анализа тезауруса заголовков докладов.

Отличительная особенность собрания текстов из цифровой библиотеки сайта музея, посвященного культуре микстекского региона – это отсутствие практики составления рефератов статей или

разделов книг, кроме того, заголовки статей и разделов книг здесь особенно кратки и, можно сказать, интригующие, но малоинформативные. В этих условиях поиск в Интернете текстовых документов по культуре Микстеки особенно затруднен. В то же время, публикация документов в полном объеме на сайте музея для анонимного пользователя признана авторами материалов шагом нежелательным и преждевременным, поскольку по их мнению это не позволит распродать тираж. Следует заметить, что цены на эти книги достаточно высоки.

Поэтому было предложено заменить заголовки новым *поисковым образом документа* – совокупностью слов в тексте из проблемно-ориентированного словаря, составленного по всем текстам библиотеки. Для составления такого словаря нет необходимости иметь текстовый документ в виде, пригодном для чтения. Достаточно иметь полный набор слов текста, представленный, например в алфавитном порядке, что нетрудно сделать в редакторе текстов. На это компромиссное решение авторы текстов согласились и полученного от них набора текстов оказалось достаточно для составления первого варианта проблемно-ориентированного словаря.

Для экспериментов выбраны сто текстов по микстекской культуре. Это тексты сборников трудов, подготовленных специально для «Недели микстекской культуры», проводимой с 2001 года в Микстекском Технологическом Университете (город Уахуапан де Леон, штат Оахака), тексты, опубликованные на разных сайтах в интернете, тексты переданные разработчикам авторами. Часть материалов была передана разработчикам сайта правительственной организацией «Culturas populares», с правом распространения через сайт (портал).

Небольшой объем выборки определяется необходимостью привлечения экспертов для отдельных оценок и экспериментов, а на больших коллекциях документов (хотя для кого-то это может показаться странным) оценки экспертов оказались неустойчивыми.

Для составления поисково-ориентированных словарей на испанском языке использовалось «домашнее» программное обеспечение, которое в значительной степени пересекается с демоверсиями системы ИРБИС, поэтому проблемно ориентированные словари лежат в кодировке windows, которая не учитывает особенностей орфографии испанского и, вообще говоря, микстекского языков. Диграфы **LL**, **QU**, **GU** и **RR** не вызывают трудностей в выделении. Однако в испанском языке над гласными (**A**, **E**, **I**, **O** и **U**) может писаться ударение для обозначения ударного слога (**á**, **é**, **í**, **ó**, **ú**), используется тильда над **Ñ** и трема над **U** (**ü**) для указания на раздельное прочтение. Поскольку ранее разработчики системы ИРБИС имели дело только с русскими и английскими текстами, эти особенности испанской орфографии не были учтены и в данное время мы имеем в проблемно ориентированном словаре только испанские слова, не содержащие слов с ударениями, тильдой и тремой. Таким образом, словари оказались искусственно обеднены.

В словарь по необходимости попали слова на микстекском языке, однако у различных авторов в их текстах и у пользователей (при написании ими запроса к поисковой машине) транскрипция микстекских слов не всегда совпадает. Например, самоназвание этнической группы миктеков в научных текстах имеет написание **ÑUU SAVI**, **ÑUU DZAVUI**, а в учебнике для начальной школы – **ÑUU SABI**. Нет оснований утверждать, что число таких разночтений обозримо, поэтому на данном этапе приходится в качестве средства локального поиска непосредственно на самом библиотечном сайте предлагать выбирать поисковые слова из заданного набора проблемно ориентированных слов.

Составление проблемно-ориентированного словаря и поисковых образов документов проведено по методике, описанной в работе [4]. Поисковые образы в этом случае представляют собой совокупности слов, встреченных в документе и в предметно – ориентированном словаре.

Одной из форм представления поисковых образов текстовых документов является матрица типа «текст-слово». Это представление в виде таблицы Excel удобно для поиска текстов по предметно-ориентированному словарю с использованием PHP.

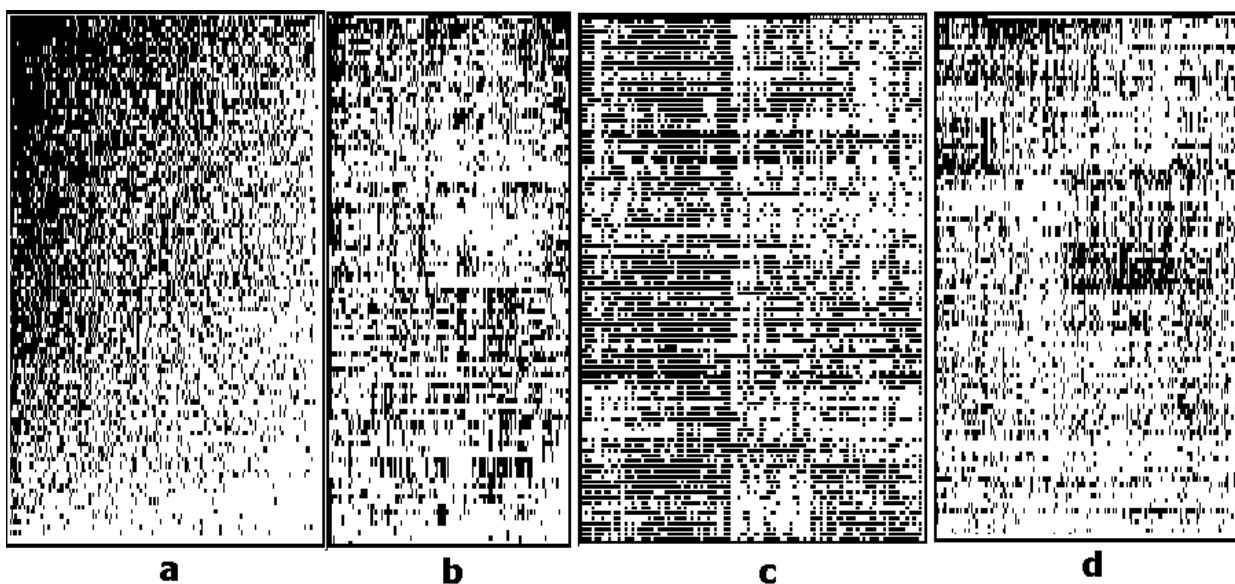


Рис. 1. Тоновые портреты матриц для 100 текстов в разных словарях

На рисунке *1a* представлен тоновой портрет матрицы типа «текст-слово», каждая строка которой фактически представляют собой поисковый образ документа в пространстве предметно ориентированного словаря. Столбцы упорядочены по убыванию применения слова во всех текстах, строки – в порядке убывания числа слов в каждом тексте. Структура матрицы не обнаруживает существенного деления на кластеры как текстов, так и слов. На рисунке *1b* видно, что после удаления высокочастотной части словаря возникает блочная (кластерная) структура.

На рисунке *1c* показан тоновой портрет матрицы тех же текстов в пространстве слов из словаря запросов. Слова, которые не были встречены ни в одном тексте, не показаны. Но эти слова отражают «социальный заказ» пользователей и будут использованы для наполнения текстовой базы новыми документами. Очевидно также, что в словаре запросов выделяется высокочастотная часть – это слова, встречающиеся в большинстве текстов, и низкочастотная часть – слова, встречающиеся в малом числе текстов. Очевидно также, что при желании возможно выявить блочную структуру матрицы – отдельным кластерам слов соответствуют специальные кластеры текстов.

В данном случае особенности испанской орфографии влияли в меньшей степени, чем при составлении предметно-ориентированного словаря, однако различия в транскрипции микстекских слов и отсутствие испанской клавиатуры у пользователей из США также сказалось на качестве словаря. В частности, в запросе к поисковой машине многие пользователи игнорировали проставление ударения. Это приходится учитывать при формировании принудительного меню – фиксированного словаря поиска в текстовой базе.

Состав предметно-ориентированного словаря остается объектом экспериментов. На рисунке *1d* представлен тоновой портрет матрицы текстов в объединенном пространстве низкочастотной части слов исходного предметно-ориентированного словаря и низкочастотной части из словаря запросов. Здесь также видна блочная структура, что означает возможность кластеризации текстов и слов и выявления соответствия кластеров слов кластерам текстов. Последнее значительно облегчает процедуры поиска в электронной библиотеке: на запрос выдается кластер текстов, соответствующий кластеру слов, которому принадлежит поисковый термин.

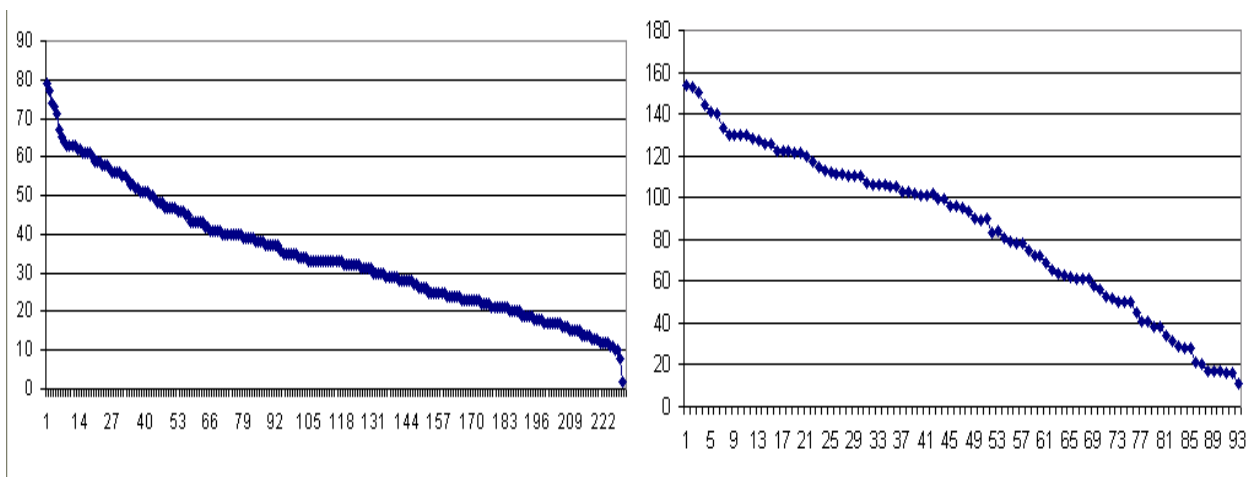


Рис. 2. (пояснения в тексте)

На рисунке *2a 2b* представлены соответственно распределение числа текстов, в которых встречены 232 слова из проблемно ориентированного словаря и число слов из словаря в каждом из текстов. Эти распределения соответствуют матрице на рисунке *1a*. Каждое слово содержится в среднем в 35% поисковых образов, каждый поисковый образ содержит в среднем 39% слов из проблемно ориентированного словаря. Для тех пользователей, кому эти объемы кажутся обременительными предлагается сокращенный словарь из левой половины матрицы «текст-слово» (рисунок *1b*).

3. Релевантность результатов при использовании поискового образа документа

Ранее [5] нами была проведена оценка возможности совпадения результатов кластеризации статей и их рефератов на материалах конференций:

- International Federation of Classification Societies 2000 по 113 документам с проблемно-ориентированными словарями в 600 и 200 слов и
- CICLing 2002 (Conference on Intelligent Text Processing and Computational Linguistics; www.CICLing.org) по 47 документам со словарями в 130 и 70 слов.

В результате автоматического и экспертного разбиения документов на 5-7 кластеров совпало от 50 до 60 % оценок принадлежности рефератов тем же кластерам. В нашем случае следует ожидать лучшего совпадения ожидаемого содержания документа с его поисковыми образами.

Наш случай, связанный с отсутствием рефератов осложняется еще и тем, что в пространстве предметно-ориентированного словаря не удастся выделить несколько сильных и кластеров с большим числом элементов и в которых межкластерное расстояние было бы существенно больше расстояний между элементами одного кластера.

Однако там же [5] было показано, что реферат статьи является текстом с бедным по содержанию (poor knowledge text), в котором встречается, как правило не более двух значимых для пользователя проблемно-ориентированных слов. В нашем случае поисковый образ документа при словаре в 200-300 единиц содержит около 90 значимых слов.

Для оценки погрешности идентификации документа по его поисковому образу невозможно применять кластерный анализ документов и их поисковых образов. Если не менять проблемно-ориентированный словарь, они просто совпадают и поэтому предлагаемый поисковый образ документа функционален, он обеспечивает релевантность поиска. Неявным образом это признается и используется всеми, кто применяет аппарат кластерного анализа текстов в АРМ «Администратор» системы ИРБИС.

С другой стороны возможно провести экспертное сопоставления документа без заголовка и его поискового образа. Для этого требуется привлечь достаточное количество экспертов, что трудно сделать в условиях дефицита специалистов по проблемам микстекстской культуры. Поэтому данная экспертиза пока не завершена.

Пользователю электронной библиотеки предоставляется право выбора словаря: из тесктов запросов, пришедших на сайт и/или из документов, лежащих в базе или в уже выделенной части базы. На сайте пользователь может найти описание упрощенных процедур поиска, прав пользователя и автора документа.

4. Заключение

В связи с отсутствием в регионе традиции составления рефератов статей поисковые образы в совокупности с заголовками документов могут быть использованы как релевантные представители этих документов, пригодные для замещения ими рефератов в полнотекстовых базах и в поисковых системах, когда доступ к самим документам ограничен.. Соответствие поисковых образов документов, их заголовков и текстов самих документов (статей и разделов книг) позволяет вести поиск в интернете текстовых документов на сайте более эффективно по сравнению с поиском только по заголовкам статей.

Литература

1. Makagonov, P., Reyes C., «Modelo conceptual para el diseño e implementación del sitio Web de un museo regional», Revista TEMAS de Ciencia y Tecnología, Vol. 11, número 31, Enero-Abril 2007, pp. 71-75
2. Pavel Makagonov, Celia Bertha Reyes Espinoza. Elements and Principal Stages in the Design of Non-Profit Websites, Proceedings 2nd WSEAS International Conference on Computing Engineering and Applications (CEA'08) pp. 115-119.
3. Библиографический справочник Трудов Вольного Экономического Общества. Том IV (1983-2000) Отв. Редактор М. М. Загоруйко. Москва, 2000., ISBN 5-94160-005-4, с. 766
4. Makagonov, P., Sboyachakov, K., Classification of Abstracts for Decision-making at the Conference. In: Proceedings of 7-th Intern. Conf. IFCS-2000 (International Federation of Classification Societies), Namur, Belgium, p. 124.
5. Pavel Makagonov, Mikhail Alexandrov, Alexander Gelbukh. *Clustering Abstracts instead of Full Texts*. TSD-2004: 7th International Workshop on Text, Speech and Dialogue, Brno, Czech Republic, September, 2004. Lecture Notes in Artificial Intelligence (indexed by SCIE), N 3206, ISSN 0302-9743, ISBN 3-540-23049-1, Springer-Verlag, 2004, pp. 129-135.