ISSN 0121-1129
eISSN 2357-5328

Towards a supervised rescoring system for unstructured data bases used to build specialized dictionaries

Hacia un sistema de ponderación supervisado de bases de datos no estructuradas utilizadas en la construcción de diccionarios especializados

Rumo a um sistema de ponderação supervisado de bases de dados não estruturadas utilizadas na construção de dicionários especializados

Fecha de Recepción: 18 de Octubre de 2014
Fecha de Aceptación: 28 de Noviembre de 2014

Antonio Rico-Sulayes*

Abstract

This article proposes the architecture for a system that uses previously learned weights to sort query results from unstructured data bases when building specialized dictionaries. A common resource in the construction of dictionaries, unstructured data bases have been especially useful in providing information about lexical items frequencies and examples in use. However, when building specialized dictionaries, whose selection of lexical items does not rely on frequency, the use of these data bases gets restricted to a simple provider of examples. Even in this task, the information unstructured data bases provide may not be very useful when looking for specialized uses of lexical items with various meanings and very long lists of results. In the face of this problem, long lists of hits can be rescored based on a supervised learning model that relies on previously helpful results. The allocation of a vast set of high quality training data for this rescoring system is reported here. Finally, the architecture of sucha system,an unprecedented tool in specialized lexicography, is proposed.

Keywords: unstructured data bases, supervised rescoring, specialized lexicography, dictionary making

* Ph. D. Universidad de las Américas Puebla (Cholula-Puebla, México). antonio.rico@udlap.mx

Resumen

El artículo propone la arquitectura de un sistema que usa valores previamente aprendidos para reordenar resultados de búsquedas en bases de datos no estructuradas al construir diccionarios especializados. Un recurso común en la construcción de diccionarios, las bases de datos no estructuradas han sido útiles ya que proveen información sobre unidades léxicas, tal como la frecuencia o ejemplos de uso de las mismas. Sin embargo, en la construcción de diccionarios especializados, cuya selección de elementos léxicos no depende de la frecuencia, el uso de estas bases de datos queda restringido a la simple ejemplificación. Incluso en esta tarea, la información de las bases de datos no estructuradas puede no ser muy útil si se buscan unidades léxicas con un uso especializado pero con varios otros significados que producen largas listas de resultados. Ante este problema, estas listas pueden ser ponderadas usando un modelo de aprendizaje automático supervisado que se apoye de los resultados previamente útiles. La recolección de un vasto conjunto de datos de alta calidad para este sistema de ponderación es reportada aquí. Finalmente, se propone la arquitectura de tal sistema, el cual representa una herramienta sin precedentes en la lexicografía especializada.

Palabras clave: bases de datos no estructuradas, listas de hipótesis supervisadas, lexicografía especializada, construcción de diccionarios.

Resumo

O artigo propõe a arquitetura de um sistema que usa valores previamente aprendidos para reordenar resultados de buscas em bases de dados não estruturadas ao construir dicionários especializados. Um recurso comum na construção de dicionários, as bases de dados não estruturadas têm sido úteis já que fornecem informação sobre unidades léxicas, tal como a frequência ou exemplos de uso das mesmas. Porém, na construção de dicionários especializados, cuja seleção de elementos léxicos não depende da frequência, o uso destas bases de dados fica restringido à simples exemplificação. Incluso nesta tarefa, a informação das bases de dados não estruturadas pode não ser muito útil se são procuradas unidades léxicas com um uso especializado, mas com vários outros significados que produzem longas listas de resultados. Perante este problema, estas listas podem ser ponderadas usando um modelo de aprendizagem automática supervisada que se apoie nos resultados previamente úteis. A coleta de um vasto conjunto de dados de alta qualidade para este sistema de ponderação é reportada aqui. Finalmente, se propõe a arquitetura de tal sistema, o qual representa uma ferramenta sem precedentes na lexicografia especializada.

Palavras chave: bases de dados não estruturadas, listas de hipótese supervisadas, lexicografia especializada, construção de dicionários.

I. Introduction

The final goal of this article is describing a route to build a system that reorganizes the results given by unstructured data bases using information about previously helpful hits. The context where such a system is being proposed is the construction of a dictionary, specifically of a substandard language dictionary. This kind of dictionary aims at describing the vocabulary of a specialized domain which covers various language uses, including colloquial or relaxed interactions, communication in popular or lower socioeconomic contexts, and stigmatized or rude forms of expression [1, 2]. Given the diverse situations where substandard language is used, the use of frequencies or other simple distributional information is not very helpful to identify and work with this kind of vocabulary in large unstructured data bases. Therefore, to maximize the benefit of using unstructured data bases, also known as textual databases [3] or linguistic corpora [4], a novel approach is needed. The approach here proposed is derived from two traditional steps in dictionary making, which include gathering all previous related lexicographic work and looking for new materials to offer an added value in the dictionary derived from them. However, the new materials here collected will have a two-fold contribution, as they will be also used to train a supervised rescoring system that improves the subsequent interaction with unstructured data bases. This article describes a proposal to build such a system, which has the potential to become a strong contribution to specialized dictionary making.

II. Unstructured data bases and dictionary making

In the construction of a substandard language dictionary for Mexican Spanish, the preliminary results of using three unstructured data bases are presented here. The idea of using natural language unstructured data bases to build dictionaries is almost as old as the idea of creating this kind of data bases for language studies [5]. While the oldest textual or unstructured data base created for linguistic applications, the Brown Corpus, dates back to 1961 [6], there have been projects to build dictionaries using this type of data bases since 1969 [5]. When unstructured data bases were first introduced in lexicography, the discipline that studies dictionary making [7], they were exploited in the construction of general dictionaries. These dictionaries attempt to describe the entire lexicon used by the speakers of a given language with an emphasis in frequent words and meanings [8]. The first project designed to build a general dictionary in its entirety using an unstructured data base was the Collins Cobuild English Language Dictionary [5]. The first edition of this dictionary appeared in 1987, with a second edition in 1995. In recent years, the use of unstructured data bases has been extended in lexicography to specialized dictionaries (which only cover a section of the lexicon of a language [8]). This extension has been particularly prolific in the English language. A good example of this is the Collins Cobuild project, formerly referred to as the pioneer work in general dictionary production [5]. Regarding specialized dictionaries, this project has produced a whole suite of didactic dictionaries. This type of dictionaries are aimed at not only helping users find word meanings but helping them use words in sentences and solve practical problems with them [9]. In order to give just a few recent examples, the didactic dictionaries resulting from the Collins Cobuild project include a number of school dictionaries --targeted to particular groups of students [10]--, such as elementary school students [11], upper-intermediate and advanced learners of English [12], and both students and teachers [13]. All these very recent dictionaries, published between 2013 and 2014, are derived completely from an unstructured "4.5-billion-word data base of the English Language" [12].

In contrast to the prolific use of unstructured data bases in English, dictionary-making projects completely supported by unstructured data bases are both, more recent and less prolific in the Spanish language. Regarding general dictionaries, there are only two recent projects that have used unstructured data bases to guide the entire construction of their dictionaries [14, 15]. It should be noted that these two dictionaries are integral dictionaries, a specific form of general dictionaries. As the latter ones, integral dictionaries attempt to cover all frequent words in some language [8], but they specifically target a language as used in a given country [16]. Be sides these two examples, there is one more lexicographic project in Spanish entirely guided by an unstructured data base. This project has produced two [17,18] specialized dictionaries of collocations, which are multi-word combinations that appear frequently in the language [6]. It should also

be mentioned that the second dictionary [18] of the two just listed is a concise version of the first one. Following this last comment, it is also worth noting that the first integral dictionary listed above [14] has a number of related works. As the final result of a four-decade project that began with the construction of an unstructured data base, this dictionary produced three preliminary versions [19-21]. Therefore, with a total of three dictionary-making projects, two for general dictionaries and another for specialized ones, the list of projects completely supported by unstructured data bases in Spanish is rather short. Chronologically speaking, this kind of project is also more recent in Spanish than in English. Although there is a dictionary in Spanish [21] as old as the first English dictionary above mentioned [5], the latter is a full-fledged product closer at least in its goals to the two integral dictionaries in Spanish [14, 15], which appeared more than twenty years later.

A. Applications of unstructured data bases in lexicography

As to the concrete use of unstructured data bases in lexicography, they have two well-known applications in the construction of general dictionaries. First, the data base can be a source of frequencies and other statistical information used in the selection of headwords, which are the words or lexical items for which entries are compiled in a dictionary [22]. An example in Spanish of this use is [14]. Second, the data base can be employed to find lists of examples in use for specific words; these lists are called key words in context or concordances in lexicography [23]. This use of unstructured data bases is aimed at identifying words meanings and other linguistic information. This was the use of unstructured data bases in [24], to give another example in Spanish. The projects that use unstructured data bases for the first application, obtaining frequencies to design their headword list, often use them for the second application too, finding examples and other linguistic information. This was actually the full use of unstructured data bases in [14]. It is possible to say, then, that these projects are completely supported by unstructured data bases, as the three Spanish language projects described in the former paragraph.

In the case of specialized dictionaries, using frequencies in unstructured data bases to determine what words to include in the dictionary is not feasible. This is because to know the frequencies of specialized vocabulary items requires knowing previously which these words are. The situation represents a chicken-and-egg problem. In order to get specialized vocabulary it is necessary to get vocabulary items frequencies, but getting these items frequencies requires knowing the vocabulary. The core of the problem is that frequency alone is not correlated to specialized domains of a language. An alternative approach to apply unstructured data bases for vocabulary selection in specialized dictionaries is to label documents with tags related to specialized language domains [23]. Using thes etags, the vocabulary in domain specific documents can be processed to obtain a wordlist in such a domain. The issue with this method is that the resulting wordlist has to be filtered to single out domain specific vocabulary. Even if the word list is filtered automatically by comparing it with a general vocabulary list, in order to get a high quality list of domain specific vocabulary, the list has to be sent eventually to specialists. These specialists can then select items belonging exclusively to a specialized form of language. In a more automatic approach, the tasks of entity recognition [25] and terminology extraction [26] have been helpful in finding words belonging to particular domains in large repositories of unstructured data. However, entity recognition is rather oriented to identify people, organizations and location names, as well as numeric expressions such as dates, times, money, and percentages [25]. Therefore, this task is not particularly relevant for a substandard language dictionary project. Terminology extraction, on the other hand, is also dependent on the previous identification of concepts that are central to a domain [26]. Taking all this into account, an unsupervised approach for the automatic recognition of substandard lexical items, as this language domain has been defined at the beginning of this article, is not practical in the construction of a high quality dictionary.

The second application of unstructured data bases to the construction of dictionaries, identifying meanings and other linguistic information of previously selected words, has also become popular in lexicography, as in the general dictionary [24]. In specialized lexicography, this approach has also been supported by the construction of specialized unstructured data

bases, for which there have been projects since the mid-eighties [27]. The drawback of this approach is that building ad hoc unstructured data bases requires a great amount of time and it still requires consulting specialists to select domain specific vocabulary.

An option to bring these data bases into specialized lexicography is using large, general language databases already available and collecting a preliminary list of words from secondary data. This type of data source not only is standard in lexicography [8,1], but has been widely successful in social sciences in general [28]. If secondary sources can provide cheap data in the form of a prolific wordlist, this list can then be used to gather new high quality, domain-specific, unstructureddata. This unstructured data would have a three-fold contribution. First, it would confirm the existence of secondary data in spontaneous language databases, eliminating the drawbacks of gathering secondary data in lexicographic work [29, 27], such as including obsolete vocabulary, recycling mistakes, or missing new information. Second, it would provide examples for the construction of the new dictionary – this natural language examples offer a number of advantages for the dictionary user and are generally praised in the literature [30,31]. Finally, the most important application in this article is the use of this type of data to train an automatic system to speed up the work when interacting again with unstructured data bases. This would make the construction of specialized dictionaries, through the use of unstructured data bases, a progressively improved cycle.

The rest of this article describes how a fair amount of human resources have been allocated to collect a large preliminary list of words from secondary data. The items in this list have been manually searched in three unstructured data bases and the results have been fed into a relational data base. With this documentation process, a fair amount of new unstructured data has been collected. Using all these data to train the supervised rescoring system, whose architecture is proposed in the last section of this article, seems rightly feasible. If the resulting system is successful in improving the search of new lexical items in unstructured data bases, it would be an unprecedented tool and a strong contribution to specialized dictionary making.

III. Collecting data for a supervised rescoring system

The section above has described a route to build a supervised rescoring system that speeds up the use of unstructured data bases in specialized dictionary making. Along the construction of the rescoring system, the steps described in the route will also update secondary sources and gather headword examples. The first step in the route is the collection of secondary data to search their lexical items in unstructured data bases. Being part of an actual project to build a substandard language dictionary, this preliminary step was implemented here in two stages. This first stage took place in 2003-2004 and was followed by a statistical analysis of the validity of secondary data and its representation in unstructured data bases. In the second stage, conducted throughout 2014, the collection of secondary data was updated and completed. In the second step of the route, a large number of collected lexical items have been searched in three data bases to gather training data and collect information for the entries that will be part of the dictionary. The rest of this section describes the results obtained in these two steps.

A. Extraction and validation of secondary data in a dictionary project

In the academic year of 2003-2004, a group of students at Instituto Tecnológico de Monterrey, Campus Puebla, extracted headwords from all secondary lexicographic materials that included substandard lexical items it the preceding decade. These secondary sources of data – previous dictionaries, vocabularies, or glossaries [8,1] – consisted of seven identified works, here listed in descending chronological order [32-34, 20, 35, 24, 36]. With a group of eight students who worked 60 hours one semester and another group of ten students who worked up to 50 hours the following semester, a word list of 13,349 lexical items belonging to Mexican Spanish substandard vocabulary was collected. Since the lexical items in the list came from sources of a diverse quality, a first important research question was whether the lexical items in this list were actually used by Mexican Spanish speakers and whether this was reflected in an unstructured data base, aimed at representing in general this dialect of

Spanish. The opposite and complementary research question whether this kind of unstructured data base represented a minimal, but substantial part of this list was also raised. These questions were answered with a strict statistical approach described in detail in [37] and briefly presented below.

For a rigorous statistical analysis, a representative stratified random sampling of the headwords extracted from each secondary data source was made. Whereas some dictionaries had contributed with a large list of headwords relevant to the study, such as [34] with 9,613 lexical items, other rendered proportionally smaller lists, like [33] and [36], with 141 and 34 lexical items respectively. In order to avoid a bias in the estimation of the volume of materials after attempting to document all the items in the wordlist, the size of each dictionary contribution had to be considered. Following [38], a representative sample was selected randomly from each dictionary. The entire representative stratified random sample had a total 1,347 lexical items. The ensuing documentation process was aimed at getting a minimum of one example and maximum of two per lexical item. These lexical items were searched in the first publicly available unstructured data base of Spanish [39], the only at the time that also allowed queries by country [37]. This unstructured data base has over 160 million words, with approximately 40% of those materials from Mexican origin. By searching only in the Mexican section of the data base, the items in the sample rendered a total 1,138 examples for 645 lexical entries. This number was calculated after subtracting the number of lexical entries with the same meaning in more than one dictionary. Extrapolating the results to the whole list, it was estimated that its documentation should render 8,931 examples for 5,192 lexical entries. These expected results were compared to figures in substandard dictionaries for other language contexts, such as the European Spanish work [40] (with 5,662 lexical entries and no examples), theAmerican English dictionary [41] (with 1,758 lexical entries and 2,232 examples), and the general English dedicated volume [42] (with 7,626 lexical entries and 6,093 samples). As it can be easily appreciated the estimated results of the projected dictionary seemed promising and competitive. The completion of the project, however, was not undertaken due to the lack of financial support.

Ten years later in 2014, the second stage of the secondary data collection has been conducted at Universidad Autónoma de Baja California. During this stage, the extraction of lexical items has been extended to include all secondary data that appeared in the last decade. With this extension, the resulting list covers twenty years of substandard Mexican Spanish lexicography. This list currently includes 36,432 lexical items, and has been extracted from 14 different lexicographic works. These works include, besides the seven dictionaries formerly referred to, the following seven more listed in descending chronological order [14, 19, 43 - 47]. With the prolific results of this second stage, the need to automate or assist by computational means the dictionary-making process became apparent.

B. New training data from unstructured data bases

Having in mind the need to speed up the work lying ahead, a search of lexical items in unstructured data bases was conducted. In this respect, new unstructured data bases for general Mexican Spanish became available. Therefore, the project was extended to include two more data bases. One of these is the Mexican only unstructured data base [48] with about 2 million words. The other was the historical data base for general Spanish [49], at least the part of its materials produced after 1921. As to the size of this last data base, it has 250 millions words, with a small portion of them, 26%,dedicated to all Latin American Spanish, and only half of this portion covering the last three centuries. These two unstructured data bases were included with the same searching criteria of the formerly commented corpus for general Spanish [39].

The second step in the route to build a rescoring system was collecting new unstructured data in the form of examples for lexical items in secondary sources. This collection of data has currently reached 9,282 examples for 3,462 lexical items. These results have been obtained after searching only 72% of the wordlist collected in the first stage of step one. For the accomplishment of these results, the accumulated work of a 14 student team currently adds to more than 1,800 hours of service. The amount of materials so far allocated, not only confirms the prediction made by the statistical analysis conducted in 2004, which estimated

gathering 8,931 examples, but it also represents a solid training data set for the supervised rescoring system proposed in the next section.

IV. A system architecture proposal to improve specialized dictionary building

The former section has described the investment of a considerable amount of human work collecting materials to build a specialized dictionary. It has also been mentioned that this work is far from being completed. In the mere validation of secondary data, just a fourth of the collected materials have been searched in unstructured data bases. If a second round of queries was to be tried, in an attempt to improve the results obtained by the first documenting team, the work currently performed would represent slightly over a tenth of all the interaction with unstructured data bases expected in this project. One of the main problems is the search of words with several meanings that render long lists of concordances when queried in large data bases. Besides how little has been done in the documenting process, at the end of it just raw materials will have been collected and a great part of the dictionary-making process will be still pending. If this is true after investing near 3,000 hours of human work, something should be done to speed up the process and make it more efficient. In order to respond to this problem, this article proposes the construction of a supervised rescoring system. The architecture is not particularly complex and the most taxing aspect of it, which is the gathering of high quality training data, has already be done in this project. Despite its straightforward design, the proposed system is unprecedented in specialized dictionary making and if successful, it should be a clear contribution to this field.

The leftmost part of Fig. 1, which describes the architecture of the rescoring system proposed, shows the steps that have already been performed in the data collection steps. Performed by individuals, the first two tasks, *Headword extraction* and *Update Headword List*, were undertaken during the two stages of the first step described in the former section. The next two tasks, *Exemplify headword* and a manual *Query*, represent the second step described above. The new part in the work pipeline consists of the next two automated steps.

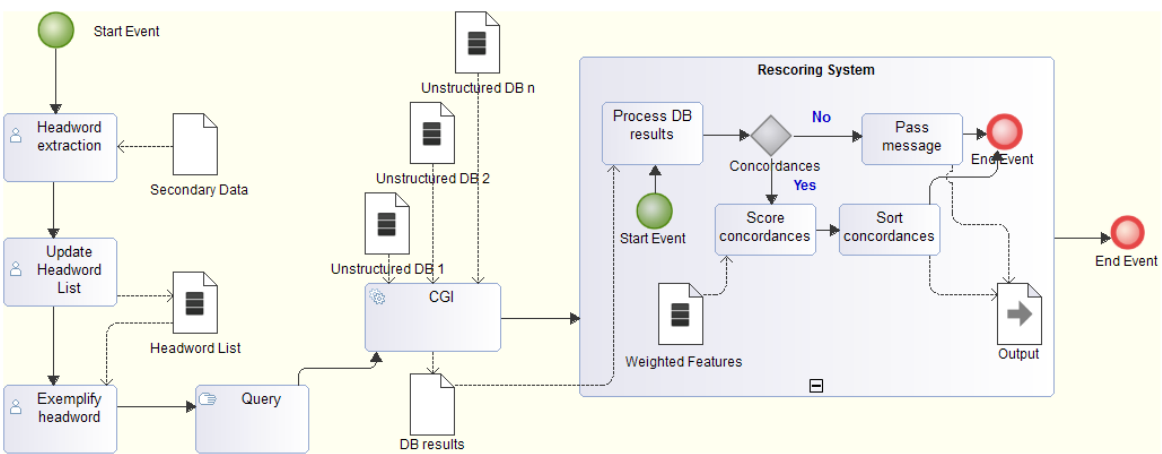

Fig. 1. A supervised rescoring system for unstructured data in specialized lexicography

The first new step in Fig. 1 is a simple *GCI* service that adapts the manual queries by exemplifiers to the interfaces of different unstructured data bases and harvests the results of those queries. Then a sub-process contains the *Rescoring System* itself. In a first task within this sub-process, the results of data bases are read. If nothing has been retrieved, the no-hits result is passed up to the user (an action expressed in the *Pass message* task) and the sub-process is terminated. If concordances have been obtained they are evaluated about their relevance for the exemplification task. As formerly commented,

the dictionary project described in this article has collected over nine thousand concordances that successfully rendered examples for substandard lexical entries. The words in these examples will be processed to obtain a list of differentiated word forms, or types, and their frequencies. Document level tags, known in information retrieval as metadata [50], such as the author and title of documents, will be identified and counted to obtain their frequency. All elements in the resulting list of word types will then be evaluated individually regarding how relevant they are to the definition of a class "substandard language". In information retrieval, there are a number of feature evaluation techniques that can be used for this purpose, such as information gain, chi square, or correlation-based feature subset selection. All of these techniques are available in the machine learning workbench, Weka [52], which is open source. In order to score the word types in the training data, they will be compared to a reference list that does not represent substandard language, but a "general language" class. This list can be easily obtained from [39]. The final list of word types with their frequencies and scores, as well as the metadata labels items, will be stored in *Weighted Features*. All the information in this data base can then be used by the task, *Score concordances*. This task will add the weights for all words in some predefined window of each concordance, to assign a global weight to it. Some fixed value for metadata of formerly observed documents in the class "substandard language" will also be added to this global weight. In a simple system, the accumulated score per concordance can be used by the next task, *Sort concordances*, to present a set of rescored results to the exemplifier.

In a more elaborate version of the task *Score concordances*, this can be a sub-process that applies one or more machine learning classification methods on some predefined window of words from concordances using the values in *Weighted Features*. Since this is a two-class problem (substandard/general language), a number of binary state-of-the-art algorithms can be employed, such as the various forms of support vector classification and Bayes Point Machines, mentioned in [52], or the binary classification methods included in [51]. Even if these sophisticated algorithms are applied, the *Score concordances* task will still depend on an accumulated score to sort concordances, making this added step ancillary to the former one. Therefore, Fig. 1 describes the whole architecture of a rescoring system that not only is feasible, given the work already done, but requires a minor investment of time or money to be implemented and tested.

V. Conclusions

This article has done a brief presentation of the practical uses of unstructured data bases in lexicography in general and in specialized dictionary making projects in particular. Given the complexity of the latter, a restricted profit from these data bases has been identified in their application to this field. At the same time, however, an alternative approach to improve the interaction with unstructured data bases in specialized lexicography has been proposed. The article has also described the intermediate steps that should permit to implement this approach collecting secondary and new data. These intermediate steps, which require a massive amount of human work, have already been performed. The validity of the work implied in these data collection steps has first been estimated with a rigorous statistical approach, and then the appropriateness of these estimations has been proved. With the results of the data collection so far done, the most important implication of the work here reported is that a large amount of data has been allocated to train the novel, supervised concordance rescoring system proposed here.

References

[1] G. Haensch, Los diccionarios del español en el umbral del siglo XX, Salamanca, Spain: Universidad de Salamanca, 1997.

[2] G. Haensch, "Tipología de las obras lexicográficas", in G. Haensch, L. Wolf, S Ettinger, and R. Werner, La lexicografía: De la lingüística teórica a la lexicografía práctica, pp. 95-187, Madrid, Spain: Gredos, 1982.

[3] S. Hockey, "Textual Databases", in J. Lawler and H. Aristar-Dry (Eds.), Using Computers in Linguistics: A Practical Guide, pp. 101-137, Routledge, 1998.

[4] P. Baker (Ed.), Contemporary Corpus Linguistics, London, UK: Continuum, 2009.

[5] S. Hockey, Electronic Texts in the Humanities: Principles and Practice, New York, NY, USA: Oxford University, 2000.

[6] H. Lindquist, Corpus Linguistics and the Description of English, Edinburgh, UK: Edinburgh University, 2009.
[7] R. A. Fontenelle (Ed.), Practical Lexicography, pp. 31-50, New York, NY, USA: Oxford University, 2008.
[8] J. A. Porto Dapena, Manual de técnica lexicográfica, Madrid, Spain: Arco libros, 2002.
[9] H. Yong and J. Peng, Bilingual Lexicography from a Communicative Perspective, Philadelphia, USA: John Benjamins, 2007.
[10] E. Bajo, Los diccionarios: Introducción a la lexicografía del español, Gijon, Spain: Trea, 2002.
[11] Collins Cobuild Primary Learner's Dictionary, (2nd ed.), London, UK: HarperCollins, 2014.
[12] Collins COBUILD Advanced Learner's Dictionary, (8th ed.), London, UK: HarperCollins, 2014.
[13] Collins COBUILD English Usage, (2nd ed.), London, UK: HarperCollins, 2013.
[14] L. F. Lara (Ed.), Diccionario del español de México, Mexico: El Colegio de México, 2010.
[15] F. Plager (Ed.), Diccionario integral del español de la Argentina, Buenos Aires: Voz Activa, 2008.
[16] R. Ávila, "¿El fin de los diccionarios diferenciales? ¿El principio de los diccionarios integrales?", Revista de Lexicografía, vol. X, pp. 7-20, 2003-2004.
[17] I. Bosque, Diccionario combinatorio del español contemporáneo: Las palabras en su contexto, Madrid: SM, 2004.
[18] I. Bosque, Diccionario combinatorio práctico del español contemporáneo: Las palabras en su contexto, Madrid: SM, 2006.
[19] L. F. Lara (Ed.), Diccionario del español usual en México, (2nd ed.), México: El Colegio de México, 2009.
[20] L. F. Lara (Ed.), Diccionario del español usual en México, Mexico: El Colegio de México, 1996.
[21] L. F. Lara (Ed.), Diccionario básico del español de México, México: El Colegio de México, 1986.
[22] B. T. Atkins, "Theoretical Lexicography and its Relation to Dictionary-Making", in R. A. Fontenelle (Ed.), Practical Lexicography, pp. 31-50, New York, NY, USA: Oxford University, 2008.
[23] B.T. Atkins and M. Rundell, The Oxford Guide to Practical Lexicography, New York, USA: Oxford University, 2008.
[24] Real Academia Española, Diccionario de la lengua española, (22nd ed.), Madrid: Espasa Calpe, 2001.
[25] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification", LingvisticaeInvestigationes, vol. 30(1), 3-26, 2007.
[26] H.F. Witschel, "Terminology extraction and automatic indexing - comparison and qualitative evaluation of methods", in Proceedings of the 8th International Conference on Terminology and Knowledge Engineering, (Copenhagen), 2005.
[27] J. Sinclair, "Lexicographic evidence" in R. Ilson (Ed.), Dictionaries, lexicography and language learning, pp. 81-94, UK: Pergamon, 1985.
[28] T. P. Vartanian, Secondary data analysis, (22nd ed.), New York, NY, USA: Oxford University, 2011.
[29] L. F. Lara, "Los diccionarios contemporáneos del español y la normatividad", in Proceedings of the II Congreso internacional de la lengua española: El español en la sociedad de la información, Valladolid, Spain, 2002.
[30] L. Bowker, "The Contribution of Corpus Linguistics to the Development of Specialised Dictionaries for Learners", in P. A. Fuertes-Olivera (Ed.), Specialised Dictionaries for Learners, pp. 155-168, Berlín, Germany: Walter de Gruyter, 2010.
[31] D. Biber, S. Conrad, and R. Reppen, Corpus linguistics: Investigating language structure and use, Cambridge, UK: Cambridge University, 1998.
[32] R. Ávila and G. Aguilar, Diccionario inicial del español en México, México: Trillas, 2003.
[33] G. Gómez de Silva, Diccionario breve de mexicanismos, México: Fondo de cultura económica, 2003.
[34] G. Colín Sánchez, Así habla la delincuencia y otros más…, México: Porrúa, 2001.
[35] A. Jiménez, Tumbaburro de la picardía mexicana: Diccionario de términos vulgares, (52nd ed.), Mexico: Diana, 1999.
[36] P. M. Usandizaga, El chingolés: Primer diccionario del lenguaje popular mexicano, (8th ed.), Mexico: Costa-Amic, 1994.

[37] A. Rico Sulayes, De vulgaridades, insultos y malsonancias: El diccionario del subestándar mexicano, Baja California, México:UABC, in press.
[38] L. R. Gay and P. W. Airasian, Educational research: Competencies for analysis and application, (7a. ed.), Englewood Cliffs, NJ, USA: Prentice Hall, 2002.
[39] Real Academia Española, Corpus de referencia del español actual, available in: http://corpus.rae.es/creanet.html, accessed: November, 2014.
[40] J. M. Iglesias, Diccionario de argot español, Madrid, Spain: Alianza, 2003.
[41] R. A. Spears, Forbidden American English: A serious compilation of taboo American English, Madrid, Spain: Alianza, 2003.
[42] J. Ayto and J. Simpson, Forbidden American English: A serious compilation of taboo American English, UK: Oxford University, 1992.
[43] J. García-Robles, Diccionario de modismos mexicanos, México: Porrúa, 2011.
[44] C. Company Company (Ed.), Diccionario de mexicanismos, México: Siglo XXI, 2010.
[45] M. P. Montes de Oca Sicilia (Ed.), El chingonario: Diccionario de uso, rehuso y abuso del chingar y sus derivados, México: Lectorum, 2010.
[46] R. Renaud (Ed.), Diccionario de hispanoamericanismos no recogidos por la Real Academia Española, Madrid: Cátedra, 2006.
[47] J. Flores y Escalante, Morralla del caló mexicano, (2nd ed.), Mexico: AMEF, 2004.
[48] El Colegio de México, Corpus del español mexicano contemporáneo, available in: http://cemc.colmex.mx/, accessed: November, 2014.
[49] Real Academia Española, Corpus Diacrónico del Español, available in: http://corpus.rae.es/cordenet.html, accessed: November, 2014.
[50] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, New York, NY, USA: Cambridge, 2008.
[51] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techiniques, (3rd ed.), Burlington, MA, USA: Morgan Kaufmann, 2011.
[52] S. I. Hill and A. Doucet, "Adapting two-class support vector classification methods to many class problems", in Proceedings of the 22nd international conference on Machine learning, (New York), pp. 313-320, ICML, 2005.