Int J Speech Technol (2017) 20:289–296
DOI 10.1007/s10772-017-9405-5

Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology

Ali Benabdallah[1] · Mohammed AlaEddine Abderrahim[1] · Mohammed El-Amine Abderrahim[2]

Received: 4 August 2016 / Accepted: 19 February 2017 / Published online: 23 March 2017

Abstract The task of building an ontology from a textual corpus starts with the conceptualization phase, which extracts ontology concepts. These concepts are linked by semantic relationships. In this paper, we describe an approach to the construction of an ontology from an Arabic textual corpus, starting first with the collection and preparation of the corpus through normalization, removing stop words and stemming; then, to extract terms of our ontology, a statistical method for extracting simple and complex terms, called "the repeated segments method" are applied. To select segments with sufficient weight we apply the weighting method term frequency–inverse document frequency (TF–IDF), and to link these terms by semantic relationships we apply an automatic method of learning linguistic markers from text. This method requires a dataset of relationship pairs, which are extracted from two external resources: an Arabic dictionary of synonyms and antonyms and the lexical database Arabic WordNet. Finally, we present the results of our experimentation using our textual corpus. The evaluation of our approach shows encouraging results in terms of recall and precision.

✉ Ali Benabdallah
benabdallah.a13@gmail.com

Mohammed AlaEddine Abderrahim
abderrahim.alaa@yahoo.fr

Mohammed El-Amine Abderrahim
med.amine.abderrahim@gmail.com

[1] Department of Computer Science, University Abou Bekr Belkaid-Tlemcen, PO box 119, 13000 Tlemcen, Algeria

[2] Department of Technology, University Abou Bekr Belkaid-Tlemcen, PO box 119, 13000 Tlemcen, Algeria

Keywords Natural language processing (NLP) · Text preprocessing · Ontology construction · Semantic relationships · Terms extraction · Linguistic markers

1 Introduction

Among the sub-domains of ontology engineering we distinguish construction of ontologies from textual documents. These ontologies are used in several areas of natural language processing such as machine translation, information retrieval, semantic annotation of resources, semantic indexing, automatic summaries of texts, translation memory, etc.

The use of texts in the ontology construction process is justified by two arguments: first, texts often carry knowledge that is stabilized and shared by communities of practice. Furthermore, even if they don't replace them completely, the texts are more readily available than the experts who lack the time to participate in the construction process (Mondary et al. 2008). But it should be noted that consultation with experts is necessary, especially in the validation step of the final ontology.

An ontology is composed of a set of concepts both hierarchically organized and structured by the relationships linking these concepts; i.e. each ontology construction process must go through two important steps: first the identification of concepts, and second, the extraction of semantic relationships.

The aim of our approach is to build an ontology from Arabic texts automatically. First, we start with the term extraction step, and then, semantic relationships are extracted from the texts based on a set of pairs of relationships, which are used as a set of examples for the automatic learning of lexical and syntactic markers.

Our paper is structured as follows: the first part focuses on our contribution and the steps required to build and methods used to prepare our corpus; in the second part we will discuss the results of applying our approach to a corpus of Arabic texts.

2 State of the art

As part of the construction of ontologies from texts, several works have been realized for the English language. Among these works, we favored operational systems that are available on the web: Text2Onto (Cimiano and Volker 2005), OntoGen (Fortuna et al. 2006), Terminae (Aussenac-Gilles et al. 2008) and others. These systems of construction of ontologies do not support Arabic text documents. Among the works on the Arabic language, we find (Mazari 2013; Zaidi-Ayad 2013; Benaissa 2012).

Mazari (2013) proposed an approach to constructing ontologies from Arabic texts using two statistical techniques for extracting information: repeated segments and co-occurrence.

Zaidi-Ayad (2013) proposed an approach for the construction of an ontology in Arabic using a statistical method for extracting simple terms, and a hybrid approach for extracting complex terms and relationships. Zaidi-Ayad (2013) applied her approach to the *Holy Quran*.

Works in the Arabic language in the field of ontology construction are very few and are not yet mature and the problem of the automatic construction of an ontology is still open for the Arabic language.

3 Contribution

To construct our ontology from a corpus of Arabic text, we adopted a process of extracting concepts and relationships from textual documents based on three steps; the first is the collection and preparation of the corpus. This step is very important because the quality of the ontology obtained will depend primarily on the quality of the processed corpus, the method of preparation and the completeness of the coverage of the domain. The second step is the extraction of simple and complex terms. To do this, we chose to use a statistical method, the method of "repeated segments" (Mazari 2013) by collecting frequently repeated words and phrases and filtering out those that represent out of domain concepts. In the third step, we use a learning method designed for linguistic markers to link the concepts extracted in the second step with semantic relationships based on the text and a set of preliminary lists of pairs for each type of relationship.

3.1 Preparation of the corpus

In the process of constructing an ontology from texts, the stage of collection and preparation of the corpus is both crucial and delicate, since the corpus is the essential source of information for the construction process (Bourigault and Aussenac-Gilles 2003).

Questions that arise in the conception of any corpus include: corpus type (a 'specialized' corpus is a corpus containing texts on a topic related to a field of knowledge, for example, in our case, "domain of computer sciences in Arabic"), suitability for the intended project, the ability to reuse this corpus, the size (number of words), representativeness (that is to say, the variety of texts, authors, sources, etc.), and the use of full texts or samples, … etc. (Marshman 2003).

After the collection of the corpus, it must be prepared for processing. This phase is performed by a set of preprocessing steps to remove some ambiguity, reduce the amount of future processing and adapt the corpus following the final objective "extraction of candidate terms" (Mazari 2013). The preparation step is divided into several sub-steps:

a. Normalization: converts the document into a moreover easily handled standard format. Before stemming, the document is normalized as follows:

   - Remove special characters and numbers, for example: ؟,),٢,٣,٤,٥,+, »…
   - Remove words and Latin characters: Latin characters are detected by their graphics.
   - Remove single letters: words of one letter in Arabic and abbreviations. For example: numbering (paragraph "B" ب "الفقرة"), date abbreviations (م: ميلادي هـ: هجري,), vowels (حروف "و"العلة), … etc (Mazari 2013).

b. Deleting of stop words: eliminate all noise words, comparing each recognized word with the elements of the list of stop words: the "stop-list". It is a list of all the tool words, link and articulation (pronouns (الضمائر المنفصلة), prepositions (حروف العطف), conjunctions (حروف الجر), etc.) for example: إلى, من, على في, أن, التي, عن, الذي, مع, بعد, بين, هذه, هذا, انه, منذ, ما, لم ... Generally, stop words that are very common (almost half of the occurrences of words of text) are not indexed because they are not informative (Vergne 2004).

c. Stemming: This is a delicate task because Arabic is an inflected language; the lack of diacritical marks (written representations of vowels) in most text written in Arabic creates ambiguity and therefore requires complex morphological rules. More, capitalized words are

not used in Arabic which makes it difficult to identify proper names, acronyms and abbreviations.

To resolve the ambiguity, Aljlayl and Frieder (2002) showed that light stemming (an approach based on the deleting of suffixes and prefixes) surpasses based root detection in the field of information retrieval. In our experimentation we have considered light stemming which is to identify whether any prefixes or suffixes were added to the word.

3.2 Extraction of candidate terms

After preparation of the corpus, we move to the step of extracting ontology concepts. First, we extract all different terms by the repeated segments method; then, we apply a filter to remove some terms that are not considered as concepts of the domain.

The method of "repeated segments" is based on the detection of text segments, consisting of pieces, appearing several times in the same text. This is a statistical technique for extracting information from texts. The repetition of running segments in a text indicates that these can be used to describe concepts of the domain corpus. The text segments may be separated by spaces or punctuation marks, and may be simple (one word) or complex (Mazari 2013).

The complex terms are identified in a window of four words in the same sentence [the number four is chosen according to the principle: a segment representing a concept contains four words maximum (Mazari 2013).

The method of "repeated segments" is based on the following proposition "a relevant term is used several times in a specialized corpus of text".

For this reason, we use "a weighting filter" to select terms with sufficient weight. The weight is measured by the weighting method term frequency–inverse document frequency (TF–IDF). It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The term frequency (TF) is simply the number of occurrences of this term in the corpus divided by the total number of terms in the corpus. The inverse document frequency (IDF) is a term-ness measure in the corpus, in the TF–IDF scheme; it aims to give more weight to the less frequent terms considered as more discriminating. It calculates the number of documents containing a target term and then divides this number by the total number of documents in the corpus. Finally, the weight TF–IDF is the product of two statistics, term frequency and inverse document frequency (Sparck Jones 1972).

If the TF–IDF is greater than a threshold: f_{min} (a threshold indicating the *relevance* of the term), the term will be kept for the next step of the construction process, otherwise it will be ignored. (F_{min} is chosen empirically, and depends on the size of the corpus).

3.3 Extraction of semantic relationships

Existing work in the field of the extraction of semantic relationships from text can be divided into two families: we usually distinguish works which are based on the frequency aspect of the corpus from which they extract related elements, and those who exploit structural clues to detect related elements, that is to say, following a symbolic approach (Claveau and Sébillot 2004).

Symbolic methods for extracting relationships are based on evidence collected in the context of an occurrence of related words to decide on their acquisition or not; the symbolic classifier is often a set of rules based on lexical clues, morphological, categorical, syntactic or others. These techniques themselves can be classified into two main families (Claveau and Sébillot 2004):

- The linguistic approaches in which structural clues given a priori (by linguistic analysis, for example) are the basic factors in the decision process.
- And approaches based on a notion of learning (markers or relationships).

Our approach is based on a method of learning markers from texts.

In the extraction methods of relationships by linguistic markers, the principle is to define initially, a set of lexical and syntactic marker lists (a list for each relationship). Next, these markers will be projected on the original corpus to identify instances of relationships.

Table 1 gives some examples of Arabic linguistic markers:

These lists allowed us to extract a number of occurrences of relationships from the corpus, but these occurrences of relationships remained insufficient relative to the size of the corpus and, also, compared to the number of concepts extracted in the previous step (several concepts remained isolated, i.e., unrelated to the ontology).

Construction of linguistic markers is then a preliminary step to identify the relationships in the corpus. Due to the specific morphology of Arabic such as vocalization and agglutination, the lists of markers for the extraction of relationships must regroup all morphological forms that may be encountered in Arabic texts. The solution is to learn these markers automatically from texts.

Table 1 List of some linguistic markers of Arabic

Relationship	Arabic linguistic markers	English translation
Hypernym and hyponym	هو نوع من، صنف من، هو، هي، هم، وغيره من أنواع،	Is a kind of, class of, is a, are, and other types of, …
Meronym	هو جزء من ، تتكون من ، تنقسم إلى ، تتألف من ،	Is a part of, consisting of, divided into, composed of, …
Antonym	ضد، عكس، بخلاف.....	Is the opposite for, unlike, …
Synonym	مرادف لـ ، يعني، مرادفه هو، له نفس معنى ،.....	Is a synonym for, means, its synonym is, has the same meaning, …
…	…	…

3.4 Marker learning

The acquisition of semantic relationships by learning extraction markers is the common point of the second type of research. Most of the time, the goal is to identify markers or clues to a semantic relationship from a set of examples in a corpus and reuse them to extract new units of relationship. It is an approach initiated by Hearst (1992) to acquire hypernym links.

The algorithm for identifying relationship pairs by learning linguistic markers from the corpus of text is:

1. For each relationship R in the set {hypernym, synonym, antonym, meronym, holonym, …} do:
2. Choose a target relation R (in our case we have tested only the tree relationships: hypernym, synonym and antonym);
3. Assemble a set of related pairs of R (extracted from a thesaurus, a knowledge base, a dictionary or built them manually by an expert);
4. Find all the sentences of the original corpus containing these pairs and record their lexical and syntactic contexts;
5. Find the commonalities between these contexts and assume that it forms an R-design;
6. Use the results found in 5 to get new pairs and return to 4.

Unlike linguistic approaches that use a priori knowledge for extracting relationships, learning approaches are based on an analysis of examples to learn extraction markers for detecting new pairs of relationships.

The set of examples containing pairs of relationships can be built manually or from an external resource such as: a thesaurus, a lexical database, dictionary or other.

The works that use external resources to extract relationships include: Girju and Moldovan (2002) who proposed a semi-automatic extraction technique of syntactic patterns using existing relationships in English WordNet.

In our approach we will use two types of external resources to obtain pairs of relationships: a dictionary, and a lexical database.

For both the synonym and the antonym relationships, we will use a dictionary (Khaled and Saad 2012) containing pairs of the best known relationships in Arabic language.

For the Hypernym relationship, we will use existing pairs in Arabic WordNet (Black et al. 2006).

4 Results and evaluation

4.1 Constitution of the corpus

In order to test our approach, we collected our corpus of text from a set of Arabic text documents containing definitions of technical terms in electronics and computer sciences in Arabic. Examples of documents in our corpus are: المعجم الموسوعي في الكمبيوتر والإلكترونيك (Encyclopedic dictionary in computer sciences and electronics), مصطلحات الحاسب والإنترنت (terms of Computer and Internet), موسوعة مصطلحات الانترنت والكمبيوتر (Encyclopedia of terms in Internet and computer sciences), etc.

These documents (in "HTML" or "PDF" format) are downloaded, selected, prepared manually (by deleting tables, charts, graphs, images …) and converted to "TXT" format.

After a manual filtering by an expert, we choose the most relevant documents and most representative of our field. The result of this operation is a corpus of 37 documents with 304,665 words and a size of 1800 KB.

4.2 External resources

To extract semantic relationships from text, our algorithm for learning linguistic markers begins with a set of examples of relationship pairs. This base can be constructed manually. In our case, we opted to take advantage of existing resources:

For two relationships "synonym" and "antonym" we chose to use a dictionary of synonyms and antonyms in Arabic; it's called: "Student's Dictionary of synonyms and opposites" (Khaled and Saad 2012) (in Arabic: "قاموس الطالب في المرادفات والأضداد") which contains about 450 pairs in total for both relationships. And for the "generalization relationship" we use a few pairs of relationships existing in the lexical database Arabic WordNet (approximately 350 pairs). Table 2 contains some examples of these relationship pairs:

4.3 Results

4.3.1 Normalization

The goal of the normalization step is to remove some special characters that do not carry information needed in later processing and thus decrease the processing time. In our experimentation, we started by detecting Arabic punctuation symbols: ["،","."," .."," …","؟","!","؛",…]. These punctuation marks are used first for segmenting the text into word sequences, and then they will be removed. Other special characters will also be deleted, for example: ["١","٢","٣","–","/","«","+","%", …]. As a result of this step, 33,515 words (11%) were deleted and 271,152 words remained of the original 304,665.

4.3.2 Deletion of stop words

In this step, we compared each word of the text with our "stop-list" which contains all the stops words of Arabic language such as: .... بعد, بين, هذه, هذا, انه, منذ, ما, لم ,إلى, من, على في, أن, التي, عن, الذي, مع If the word compared exists in the "stop-list" it will be deleted. As a result of this step, we eliminated 72,510 words (26.74%) from the words remaining after the normalization step.

4.3.3 Stemming

We removed the prefixes and suffixes from the words appearing in a predefined list. This list is stored in two files (prefix file and suffix file). After the deletion of prefixes and suffixes we found again some stop words. These stop words will be deleted using the same "stop list" of the previous step.

Examples: بعضها → بعض , أنك → أن. As a result of this step, 184,738 words remained and 13,904 words were removed (7%).

4.3.4 Extraction of "repeated-segments"

For extracting terms of our ontology we set the following parameters:

- Maximal segment size = 4 words. It indicates the maximum size of a complex term. Example: بروتوكول نقل البريد البسيط (Simple Mail Transfer Protocol)
- Weighting threshold: The weight of a term is calculated by TF–IDF. The threshold weight of a simple word is 4E−05. The threshold weight of a complex term is 5E−06. These thresholds are selected relative to the corpus size.

Our program extracts 528,692 different segments (184,738 simple terms and 343,954 complex terms), but it only selects a list of 741 segments in accordance with the thresholds defined above (552 simple terms and 189 complex terms). Table 3 shows some examples of selected segments:

Table 2 List of relationship pairs

Relationship	Hypernym		Synonym		Antonym	
Pairs	حاسوب (Computer)	آلة (Machine)	ابتكار (Innovation)	اختراع (Invention)	اقفال (Lock on)	افتتاح (Opening)
	آلة (Machine)	جهاز (Device)	ارتباط (Correlation)	اتصال (Connexion)	الإسراع (Accelerate)	التأخير (Delay)
	جهاز (Device)	شيء مصنوع (Anything made)	عداد (Counter)	حاسب (Computer)	الممتلئ (Full of)	الفارغ (Empty)
	شيء مصنوع (Anything made)	جسم (Body)	إغلاق (Close)	اقفال (Lock on)	التجميع (Assembly)	التقسيم (Division)
	برنامج حاسوب (Software)	شفرة (Code)	حاسوب (Computer)	كمبيوتر (Computer)	يئزج (Partial)	كلي (Total)
	…	…	…	…	…	…

Table 3 Examples of selected segments from the corpus

Segment	Number of occurrences in the corpus	TF	Number of documents containing the segment	IDF	TF–IDF
(Computer) كمبيوتر	3045	0.00999	27	0.72973	0.00729
(Information) معلومات	2233	0.00733	35	0.94595	0.00693
(Internet) إنترنت	5278	0.01732	14	0.37838	0.00656
(Site) موقع	1624	0.00533	15	0.40541	0.00216
…	…	…	…	…	…
(World web) الشبكة العالمية	203	0.00067	10	0.27027	0.00018
(Electronic mail) البريد الإلكتروني	101	0.00033	11	0.29730	0.00010
…	…	…	…	…	…

4.3.5 Extraction of semantic relationships

By applying the marker learning algorithm on our Arabic text corpus, our application has detected a set of *markers and pairs* for each semantic relationship. Table 4 shows some examples of these results:

4.4 Evaluation

In this section, we present the results of evaluations conducted on our textual corpus. Table 5 summarizes the results of application of our approach to the extraction of terms and relationships on a subset of words from our

Table 4 Some examples of relationship instances extracted from the corpus

Relationships	Relationship *pairs* and *markers* extracted from the corpus		
	Term 1	Detected markers (word sequences)	Term 2
Hypernym of	الملفات (Files)	[و],[غيرها],[من] (And others of)	روصلا (Pictures)
	لغات البرمجة (Programming languages)	[من],[بين],[الأصناف],[الأكثر],انتشارا], [ل] (Is among the most popular items of)	جافا (Java)
	أنظمة التشفير (Coding system)	[شكل], [من],[أشكال] (Is a form of)	يونيكود (Unicode)
	الحواسيب (Computers)	[له],[تأثير],[كبير] [على],[باقي] (Has a significant impact on the rest of)	الخادم (Server)
	…	…	…
Synonym of	البيانات (Data)	[في],[بعض],[الأحيان],[هي],[مرادفة] [ل] (Sometimes it is synonymous for)	المعطيات (Data)
	شيفرات (Codes)	[يمكن],[أن],[تسمى],[ب] (Could be called)	رموز (Symbols)
	الكمبيوتر (Computer)	[هي],[كلمة],[تعني] (Is a word that means)	الحاسوب (Computer)
	…	…	…
Antonym of	الدائمة (Permanent)	[تعني], [عكس] (Means the opposite)	ةيئاوشعلا (Temporary)
	المستفيد (Client)	[بخلاف], [كلمة] (Is opposed to the word)	الخادم (Server)
	المجاهيل (Unknown)	[على], [خلاف] (Is at odds)	المعطيات (Data)
	…	…	…

Table 5 Evaluations of ontology elements recognition results

Ontology elements	Precision	Recall	F1-measure
Terms			
Simple	0.94	0.88	0.91
Complex	0.84	0.76	0.80
Relationships			
«synonym»	0.90	0.58	0.71
«antonym»	0.91	0.67	0.77
«hypernym»	0.90	0.65	0.75
Average	0.90	0.71	0.79

text corpus. The measures used are typically the precision and recall, defined here as follows:

The precision is the number of correct entities extracted by our system divided by the total number of entities extracted by our system;

The recall is the number of correct entities extracted by our system divided by the total number of correct entities in the corpus.

The *F1-measure* corresponds to the harmonic average between precision and recall [2 × recall × Precision/(recall + precision)].

To get the recall of our extraction approach of terms and semantic relationships, we need to know how many correct terms and relationships actually exist in the test corpus. To discover this, a domain expert has read through our text corpus and manually tagged all existing correct entities. Based on the results obtained by this manual processing and the results obtained by our system, the precision and the recall were calculated as indicated in the following table:

The results presented in Table 5 concerning simple and complex terms show a precision equal to 89% on average, and a recall equal to 82% on average, which is a good level for this type of task. It may be noted in particular the high level of precision which characterizes a very significant level of reliability.

As in the case of term recognition, the validation of extracted relationships is characterized by a high precision and a low recall. The difference between precision and recall is also more accentuated in this case than for the term recognition, partly due to medium precision slightly higher but mainly by a lower recall. So we can say that the relationships produced by our system are generally good reliability but that linguistic markers automatically learned don't cover all possible forms of the Arabic language, and therefore it is necessary that our approach be applied to other texts of the same domain in order to learn more markers and thus extract more relationship instances.

5 Conclusion and perspectives

In this paper we proposed an automatic approach to ontology construction from Arabic texts. The latter is based on the extraction of terms and semantic relationships from three resources: a corpus of Arabic texts and two external resources: an Arabic dictionary of synonyms and antonyms and the lexical database Arabic WordNet. Our construction process begins with a preliminary step, which is the preparation of the corpus through normalization, removing stop words and stemming. Then, to extract candidate terms, we used a statistical method "repeated segments" followed by the application of a weighted filter. Finally, to connect new concepts extracted, we adopt a method for learning linguistic markers. The results of experimentation with our approach show good precision and recall values and they are therefore encouraging.

The work described here opens up a number of new directions for future research. First, there are several potentially valuable ways to filter extracted segments in the extraction step. Depending on the application we may want to filter out named entities, verbs, or spelled-out numbers. A second direction is to use the constructed ontology in the development of Arabic tools such as information retrieval systems or machine translation to or from Arabic.

References

Aljlayl, M., & Frieder, O. (2002). On Arabic search: Improving the retrieval effectiveness via a light stemming approach. In Proceedings of the eleventh international conference on information and knowledge management (pp. 340–347). ACM Press, New York, NY, USA, ACM DL Digital Library, http://dl.acm.org/citation.cfm?id=584848. Accessed May 1, 2016.

Aussenac-Gilles, N., Despres, S., & Szulman, S. (2008). The TERMINAE method and platform for ontology engineering from texts. In P. Buitelar & P. Cimiano (Eds.), *Bridging the gap between text and knowledge: Selected contributions to ontology learning from text* (pp. 199–223). Amsterdam: IOS Press.

Benaissa, B. (2012). Construction semi-automatique d'ontologies à partir de textes arabes, Dissertation, University of Tlemcen, Algeria.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Bertran, M., & Fellbaum, C. (2006). The Arabic WordNet Project, Proceedings of LREC 2006.

Bourigault, D., & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. ATALA (Association pour le traitement automatique des langues) http://www.atala.org/taln_archives/TALN/TALN-2003/taln-2003-tutoriel-002.pdf. Accessed December 15, 2015.

Cimiano, P., & Volker, J. (2005). Text2Onto—A framework for ontology learning and data-driven change discovery. In Natural language processing and information systems, lecture notes in computer science (pp. 257–271).

Claveau, V., & Sébillot, P. (2004). Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe, TAL (Traitement

Automatique des Langues). Vol. 45, no 1/2004, pp 153–182. http://people.irisa.fr/Vincent.Claveau/publications.html#2004. Accessed May 15, 2016.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2006). Semi-automatic data driven ontology construction system. In Proceedings of the 9th international multi-conference Information Society IS (pp. 223–226). Ljubljana, Slovenia.

Girju, R, & Moldovan, D. (2002). Text mining for causal relations. In 15th international Florida artificial intelligence research society (pp. 360–364).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th international conference on computational linguistics, (COLING'92) (pp. 539–545). Nantes, France.

Khaled, W., & Saad, D. (2012). *Student's dictionary of synonyms and opposites*. Beirut, Lebanon: Alrouqy-Verlag.

Marshman, E. (2003). Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie, In Observatoire de linguistique Sens-Texte (OLST). University of Montréal. http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf. Accessed January 15, 2016.

Mazari, A. C. (2013). Vers une approche statistique pour l'extraction des éléments de l'ontologie à partir des textes arabes. In: RML (Revue Maghrébine des langues), ISSN: 2253-0673, 8th edition, Oran Algeria (pp. 39–56).

Mondary, T., Després, S., Nazarenko, A., & Szulman, S. (2008). Construction d'ontologies à partir de textes: la phase de conceptualisation, 19èmes Journées Francophones d'Ingénierie des Connaissances, Nancy, France, LIPN—UMR 7030 University of Paris 13—CNRS.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 1, 11–21.

Vergne, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource, 7eme Journées internationales d'analyse statistique des données textuelles, GREYC—University of Caen.

Zaidi-Ayad, S. (2013). Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran), Dissertation, University of Annaba Algeria.