Modeling Semantic Change as Linked Data using Distributional Semantics: A Case on the Arabic Language

Alia O. Bahanshal
King Saud University
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia
abahanshal@kacst.edu.sa

Hend S. Al-Khalifa
King Saud University
Riyadh, Saudi Arabia
hendk@ksu.edu.sa

AbdulMalik Al-Salman
King Saud University
Riyadh, Saudi Arabia
salman@ksu.edu.s

Abstract

Semantic change focuses on the study of word usage evolution, where the new meaning of a word is somehow different from the original usage. This paper proposes a linked data model to represent semantic change identified by a distributional semantics approach applied on the Arabic language.

Keywords: Linked Data, Semantic Change, Distributional Semantics, Arabic Language

1. Introduction

Words acquire new meanings through time, and anyone who reads old texts can notice words which have different meanings today. The study of language change is essential for anyone who uses ancient literature such as religious scholars, librarians and linguists. By language change, we mean the semantic variations in language. Semantic change, also called semantic development and semantic shift, is part of the semantics (the study of meaning in a language), and it focuses on the study of word usage evolution, where the new meaning of a word is entirely different from the original usage. For example, the Arabic word "حريم" \hareem\ (women) used to mean in the old dictionary: everything forbidden, and their new meaning in the contemporary dictionary is: women. From this example, we can see how meanings changed over time. Another case in English is the word "dog" that is a specific breed of the dog, which has become later the entire race of dogs. This kind of change is called widening or extension. Semantic change occurs for political, social, economic and historical reasons or just to name things. The ability to identify semantic change would help linguists understand the evolution of words through time and recognize cultural phenomena. Furthermore, many applications could benefit from such studies, such as Natural Language Processing (NLP), automatic interpretation and translation process. For instance, to translate a recent publication, we cannot use the words meanings from an ancient dictionary since the words meanings change over time and may not reflect the current words meanings. Language change should be known to use the accurate meaning of words for the desired publication time. Furthermore, by recognizing language change, linguists would be able to identify the most used lexicons in literature and the ones diminished by observing the frequency of the words in different time periods of a corpus as in (Michel et al., 2011), and many more applications could be built.

Different approaches were used to identify words semantic change over time. One of these methods is statistical semantics, where statistics are used to determine words meanings, as stated in (Furnas et al., 1983; Weaver, 1955) that "Statistical patterns of human word usage can be used to figure out what people mean". The statistical semantics hypothesis subsumes the distributional hypothesis, which in linguistics is based on word context, as stated in (Harris, 1954) that "Words that occur in similar contexts tend to have similar meanings", and in (Firth, 1957) that "You shall know a word by the company it keeps". This hypothesis was the motivation for researchers to use distributional semantics further to measure diachronic change through time (Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Rodda et al., 2016).

Linked Data was introduced as a method to publish interlinked data forming a single global space of data from various sources (Web of Data). Its features such as openness, linking capabilities, and graph representations witnessed great attention in the field of linguistics and proved its capability in Natural Language Processing (NLP) and in representing lexical resources as open data.

The aim of this paper is to apply distributional semantics to the Arabic language to identify semantic change, where no one (to the best of our knowledge) explored this area before, to know if the methods applied to the English language would function as well. Our aim is also to propose algorithms that utilize existing methods for identifying semantic change based on Distributional Semantics Models (DSM) Vector Space Model (VSM) and Latent Semantic Analysis (LSA), where the current frameworks that used DSM relied only on visualization to identify semantic change and do not provide a broad approach to analyze and utilize the resulted visuals and information. Furthermore, we aim to propose a new model that represents the results of the distributional methods as Linked Data, and to solve existing models' lack of fundamental information needed for the semantic change identification process.

The remainder of this paper is structured as follows. Section 2 presents background information about the

Arabic language, semantic change, distributional semantics methods, Linked Data, and the *lemon* model which is used as the base for the new proposed model to represent semantic change as Linked Data. Section 3 describes related works. Section 4 introduces the dataset used in this research and the preprocessing steps. Section 5 explains the distributional semantics algorithms to identify the semantic change. Section 6 presents the proposed Arabic Semantic Change (ASC) model. Section 7 explains case studies where distributional semantics algorithms and the ASC model are applied. Finally, Section 8 provides conclusion and future works.

2. Background

In this section, background about the Arabic language, semantic change, distributional semantics, Linked Data, and the *lemon* model is presented.

2.1 Arabic Language

Arabic is the language of (Quran), Muslims religious book, and a Semitic language spoken by nearly 500 million people around the world and one of the official UN languages. Like any language, Arabic has its grammar, spelling, and pronunciation; yet it has its own characteristics which made it distinctive. Arabic is read and written from right to left (except numbers), its alphabet consist of 29 spoken letters, and 36 written characters. Classical Arabic descends Modern Standard Arabic (MSA), which is the language used in formal writing and speech, and Colloquial Arabic, which is the language spoken every day and what children speak as their first language. Arabic is written with an orthography that includes optional diacritical marks. Diacritics are extremely useful for readability and understanding, their absence in Arabic text adds another layer of lexical and morphological ambiguity. Diacritics in Arabic are optional orthographic symbols typically representing short vowels and aid the reader to disambiguate the writing or just articulate it correctly. The Quran is fully diacritized to minimize the chances of misinterpreting it. Children's educational texts, classical poetry tends to be diacritized as well. The reader should analyze the text morphologically, syntactically and semantically before reading it, i.e., restoring the diacritics. It is very rare to use diacritics in modern Arabic text. Newspapers, books, and the Internet have Arabic content that is usually written without diacritics. The Arabic language is our focus in this paper, where the semantic change is identified for it.

2.2 Semantic Change

Semantics is defined as "the study of meaning" (Lyons, 1977), "the study of meaning in language" (Hurford, 2007), "the study of meaning communicated through language" (Saeed, 1997), and "the part of linguistics that is concerned with meaning" (Löbner, 2002). Semantic development is a branch of Semantics which focus on change in the meaning of words to help researchers understand the words evolution through time (Issa and Issa, 2008). Semantic change is defined as "the gradual change in words semantics through time, where words change their meanings from one to another as a result of several life changes" and semantic development is equal to semantic change by many linguistics opinions and does not mean a rise in meaning (Qalalah, 2017). All languages in the world face semantic change from time to time and are forced by the law of change; some languages evolve faster than others in some specific time periods (Issa and Issa, 2008). The semantic change is the fastest branch of Semantics to evolve because it is bond with the human movements and life change (Abuhadeemah, 2008).

Semantic change could occur to name things. For instance, the word "انترنت" \e`ntarrnit\ does not exist in old literature but appeared as a translation of the English word Internet. The meaning of words change over time, e.g., the meaning of word "حرامي" \ḥarāmī\ in Arabic through time became the same meaning of the word "السارق" \alsāriq\, the person who made mistakes in general and became the person who steals things particularly. In the English language, the word 'mouse' that means the small animal witnessed the change of the addition of another meaning 'the computer device'. The semantic change identification using computational approach is the main goal of this paper.

2.3 Distributional Semantics

Distributional semantics is built upon distributional hypothesis, which is in linguistics is based on word context where words with similar meanings occur in a similar context (Firth, 1957; Harris, 1954). Thus, the meaning of a word is related to the distribution of words around it. The efforts to apply this hypothesis on semantics usually led to vectors and metrics, which was the motivation to investigate further vector space model (VSM) and its relationship with words meaning (Turney and Pantel, 2010). "The representation of a set of documents as vectors in a common vector space is known as the ***vector space model***" (Chowdhury, 2010)."The idea of the VSM is to represent each document in a collection as a point in space (a vector in a vector space). Points that are close together in this space are semantically similar, and points that are far apart are semantically distant" (Turney and Pantel, 2010).

The performance of information retrieval is improved when the number of vectors' components is limited. To reduce the dimensionality of vector models, Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is used, which maps documents and terms to a common conceptual space. The statistical technique used is called Singular Value Decomposition (SVD) (Golub and Reinsch, 1970). LSA creates a semantic space of a sizeable term-document matrix where terms and documents that are closely related are placed nearby each other. With SVD, the small and unrelated data are ignored, and a new space is arranged with the most associative data. VSM and LSA are further investigated in this paper to identify the semantic change in the Arabic language.

2.4 Linked Data

In 2006, Linked Data was introduced by Tim Berner-Lee as a method to publish interlinked data forming a single

global space of data from various sources (Web of Data) (Heath and Bizer, 2011). Linked Data format is understood by machines, and thus raw data can be retrieved. It is defined as: “best practices for connecting and publishing structured data on the Web” (Bizer et al., 2009). Linked Data, same as the web of documents, connects different online resources, but it interlinks both data and documents in a predefined standard format.

Resource Description Format (RDF) (Beckett, 2004) is the standard model for the expression of data and relations in Linked Data. RDF data model consists of (subject - predicate – object) triples. The subject and object could be URIs referencing to resources. The object could be a string literal, while the predicate represents the relation between a subject and an object. Linked Data was used in several domains such as Medical and Health, Education, Government, Linguistic domain and more. Data from different data sources were converted into RDF format and interlinked forming the Linking Open Data (LOD) cloud (Figure 1). The LOD cloud is a community effort founded in 2007, and the World Wide Web Consortium (W3C) assists with its production under the Linked Open Data project coordination (Bizer, 2009).

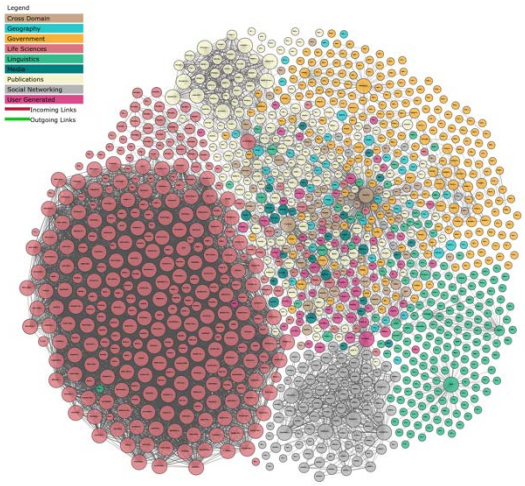

Figure 1 LOD Cloud Diagram (Abele et al., 2017)

Linked Data witnessed a considerable attention in the Linguistic domain and many tools and applications utilized Linguistic Linked Open Data (LLOD). Natural Language Processing (NLP), is one of the goals of creating a sub-cloud of datasets related to the Linguistic domain (Chiarcos et al., 2013).

In this paper, we propose an extension to LLOD that will represent the semantic change information using distributional semantics methods as Linked Data. We named the extended model "Arabic Semantic Change" (ASC).

The *lemon* model (McCrae et al., 2012a) was chosen as the base for our proposed ASC model. *lemon* model was introduced to describe vocabularies that are used to enrich ontologies vocabulary elements with information that are realized linguistically and in natural languages, and this is because ontology languages such as OWL [1] (The Web Ontology Language) and RDF lack the support of linguistic data[2]. *lemon* represents lexical entries morphological and syntactic properties and acts as a syntax-semantics interface. It was first developed within the European project “Monnet” and further developed by W3C Ontology-Lexica Community Group[2].

lemon follows “semantics by reference” principle, where the lexical meaning is stated entirely in the ontology, and the lexicon only points to the proper concept, unlike other lexical resources which include as part of the lexicon the semantic relations, such as synonymy and antonymy (McCrae et al., 2012a).

OntoLex is the new version of *lemon* core, and it was released in 2016 by W3C Ontology-Lexica Community Group[2]. Figure represents the core of *lemon* (OntoLex), which covers “the basic elements required to define lexical entries and associate them to their lexical forms as well as to concepts in the ontology representing their meaning” (McCrae et al., 2012a).

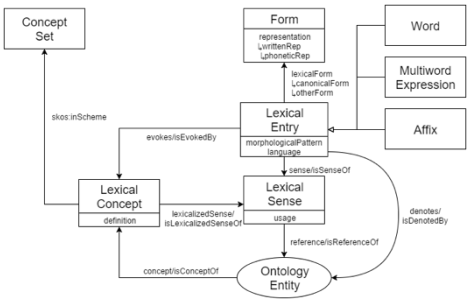

Figure 2 *lemon* OntoLex Core[2]

The main elements of the OntoLex are Lexical Entry, Forms, Semantics, Lexical Sense & Reference, and Lexical Concept. Lexical entry is the main class of the OntoLex core, and it is defined as: “A lexical entry represents a unit of analysis of the lexicon that consists of a set of forms that are grammatically related and a set of base meanings that are associated with all of these forms. Thus, a lexical entry is a word, multiword expression or affix with a single part-of-speech, morphological pattern, etymology and set of senses”[2]. There are different forms for each lexical entry from the grammatical point of view, and it is defined as: “A form represents one grammatical realization of a lexical entry”[2].

The Semantics in the model represents the meaning of a lexical entry using `denotes` property by pointing to the ontological concept following the *semantics by reference* principle. Lexical senses were introduced because the property `denotes` was not sufficient for all the linking cases of the lexical entry with the ontology. `LexicalSense` is defined as: “A lexical sense represents a reification of a pair of a uniquely determined lexical entry

[1] https://www.w3.org/OWL/

[2] https://www.w3.org/2016/05/ontolex/

and a uniquely determined ontology entity it refers to. A link between a lexical entry and an ontology entity via a Lexical Sense object implies that the lexical entry can be used to refer to the ontology entity in question."[2]. The lexical concept is defined as: "A lexical concept represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses"[2].

3. Related Work

In this section, some works related to distributional semantics methods to identify semantic change and previous models used to represent that change as Linked Data are presented.

For the distributional semantics methods, Jatowt and Duh (2014) proposed a framework for identifying semantic change at three levels: lexical, contrastive-pair and sentiment orientation levels. The framework is based on distributional semantics, where the meaning of a word is identified from contexts in which it occurs in texts. In their approach, they viewed each time period's context words as a vector and calculated the similarity between vectors. If the vectors are dissimilar, it means a semantic change has occurred. To evaluate their approach, they used visualization only and listed the top context words. In our work, we proposed an algorithm that further utilizes the top context words to identify the different meanings in the various time periods and according to these meanings the semantic change was detected. Also, we employed the method on the Arabic language to identify the semantic change.

In the area of representing semantic change as Linked Data, van Aggelen et al. (2016) proposed a model for describing semantic change and the connection with WordNet (Fellbaum, 1998) *lemon* model (McCrae et al., 2012b). The problem with this model is that it focused only on representing the similarity scores between decades and did not include any information about the meanings or the context words in different decades. Figure 3 shows the

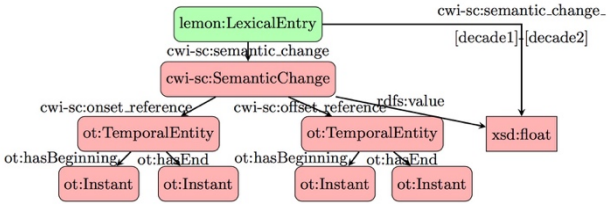

Figure 2 A model for connecting WordNet entries to cross-decade scores of lexical changes. prefix ot stands for OWL-Time and cwi-sc for the purpose-built vocabulary (van Aggelen et al., 2016)

structure of that model and the connection with *lemon*'s Lexical Entry, the old version of OntoLex[2].

The model has two time periods and their similarity scores, however, no information about the meanings are displayed. Thus, in our proposed model, this problem will be solved

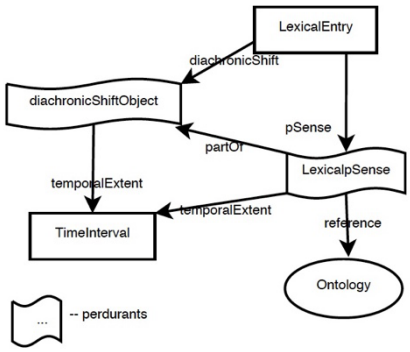

Figure 4 The lemonDia Model (Khan et al., 2014)

by including the top context words retrieved from the corpus, also we incorporated the different meanings obtained from the results of the proposed DSM algorithm.

Also, Khan et al. (2014) proposed an extension to *lemon* model, the lemonDia model to represent the diachronic change between word meanings. The problem with this model is that it focused on representing known words that changed their meanings through time and did not include information about the distributional semantics methods information such as similarity scores or context words. Figure 4 shows the structure of lemonDia model, where no information about the word's distributional semantics is contained. Therefore, in our proposed model this issue will be solved by including the similarly scores for each period along with the top context words.

4. Dataset

For the words that changed their meanings through time, N-gram data need to be collected from a corpus. This activity is required to know the most occurred terms around the words in the study and based on the distributional semantics hypothesis, the meaning of a word is known by the words around it (its context). N-gram is a sequence of n terms usually collected from text or collection of data, e.g., corpus. In this paper, the n size is five, and the 5-gram is the context words surrounding the target word, two words before and two words after, which was used by Wijaya and Yeniterzi (2011) and showed good results for semantic change identification. To collect the 5-gram, we have used KACST corpus (Al-Thubaity, 2015) a large and diverse Arabic corpus with a free access. It was developed to be used for several purpose applications, from linguistics research to developing NLP applications. The corpus size started at seven hundred million words and currently has one billion words[3] in Arabic language (Classical Arabic and MSA). KACST corpus is divided by time periods (0-600 to 2011-2020). However, the number of data in the earlier years was small in size, yet, the time period (1700-1800) is where the corpus started to have a sufficient number of data. Therefore, we chose the periods (1700-1800 to 2011-2017) in our study to ensure that a suitable amount of datasets is collected from the corpus.

After collecting the frequencies of the 5-gram words for all periods of each study word, the results were recorded in Excel sheets. Our dataset is constructed from every period,

[3] http://corpus.kacst.edu.sa

and its 5-gram unique words along with the words frequencies count. After constructing the dataset, we noticed that the collected data contained many symbols such as (“), and stop words such as the preposition word في \fī\ (in), which may affect our analysis results. Therefore, we developed a Java program to clean the data from unwanted contents (stop words, punctuations, and symbol). The program used an Arabic stop words list[4] that originally contained 750 words. This list was expanded to include 1659 items including Arabic stop words and symbols. These items were retrieved after manually cleaning the dataset and it is available for download[5].

Also, Google Books corpora[6] were used to extract 5-gram dataset for the English words to apply the proposed algorithm and model on them, and to test if they can be generalized to all languages. The Google Books corpora have 155 billion words, and the extracted 5-gram words should appear at least 40 times in the corpus, the dataset is divided by decades from the year (1810) to year (2000).

5. Using Distributional Semantics to Identify Semantic Change

Vector Space Model (VSM) is a widely used approach in Information Retrieval (IR) and NLP, and in our case, it is used to identify the semantic change. Figure 5 shows the steps needed to construct VSM to identify the semantic change.

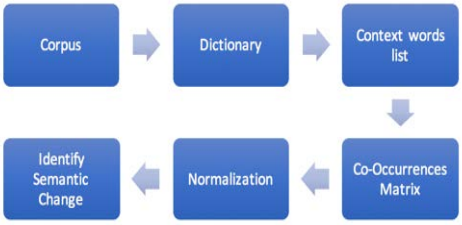

Figure 5 The steps needed to construct the VSM to identify semantic change

In Figure 5, after collecting the 5-gram (context words) from KACST corpus for all periods, they are listed as an Excel sheet. Next, we developed Arabic Semantic Change tool (Figure 6) using Java to construct a co-occurrences $M \times N$ matrix from the dataset. The matrix rows are the M context words in all positions around the target words. The matrix columns are the N time periods of the diachronic KACST corpus. In our case, the time is partitioned from the period 1700-1800 to 2011-2017.

At this point, we can view each time period in the matrix as a vector, which depends on the number of occurrences or the frequency of a context word in that period. After constructing the vectors for each time period, where the vector of a word w in a time period i is represented as $t_i[w]$, the tool calculates the weight of each vector's context word. The weight is the count of a context word w in t_i (or the term frequency (Tf) of w in t_i) divided by the total number of context words in that time period (or the

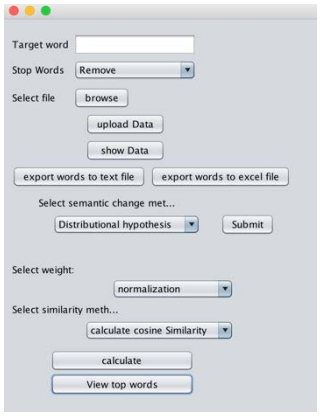

Figure 6 Arabic Semantic Change Tool

normalized length of the vector). Euclidean length is used to compute the vectors' lengths.

Afterward, the semantic change is measured by calculating the similarity between the same word vector in time period i and time period j. We denote that by $\mathrm{sim}(\mathrm{t_i}[\mathrm{w}], \mathrm{t_j}[\mathrm{w}])$. Semantic change is likely to occur if $[\mathrm{sim}(\mathrm{t_i}[\mathrm{w}], \mathrm{t_j}[\mathrm{w}]) \rightarrow 0]$. To measure the semantic change, we followed (Jatowt and Duh, 2014) approach to compute the similarity of word w vector in last time period vector similarity with all previous time periods vectors. Similarity is calculated using well-known similarity measures Cosine similarity (Huang, 2008). Next, to identify the semantic change of a word, we plot the similarity values and observe the plotted line. If the line has a steep increase or fall, it means semantic change likely occurred, and if it is almost a straight line, it indicates the word has a stable meaning through time. This result is converted into a numerical value, called the (*VSM Semantic Change Plotting Result*) and it ranges from 0 to 1 according to the similarity curve. Additionally, the top context words are retrieved to know if the meanings were changed over time. To extract the top context words, the approach by (Jatowt and Duh, 2014) was used. From each time period, the context words with less than 1% frequency are removed. Then, the frequency of each of the remaining context words a in time period t_i are then compared to the frequency of the same word in time period t_{i-1}.

$$S(\mathrm{a}, \mathrm{t_i}) = f(\mathrm{a}, \mathrm{t_i}) \Big/ f(\mathrm{a}, \mathrm{t_{i-1}})$$

To identify semantic change, the meaning of a word should be known, and the retrieved context word can be used to determine that meaning. This assumption was made after presenting the collected list of top context words to three different linguists to identify the semantic change. They gave different opinions about the rising and the reduction of the word meaning but they all agreed on that meaning could be comprehended from the list of context words in each time period. Thus, an algorithm (Algorithm 1) that utilizes this agreement was proposed to identify the semantic change.

From Algorithm 1, the $vx1$ matrix is constructed, where

[4] https://github.com/mohataher/arabic-stop-words

[5] https://github.com/abahanshal/arabic-stop-words-list

[6] https://googlebooks.byu.edu

Algorithm 1: Identify semantic change using VSM

For $i = 1$ *to* k , *where* k *is the number of time periods* d *selected from corpus* c
Do
  Retrieve v *context* words from c for each d_i
  Construct k occurrences $v \times 1$ matrix, *where rows are* v *context words and column is time period* d_i
End
For $i = 1$ *to* k
Do
Remove all context words with less than 1% frequency from d_i occurrences matrix
Compute top context words T of d_i by dividing each word a *frequency* in d_i with same word a *frequency* in d_{i-1}

$$T(a, d_i) = f(d_i, a)/f(d_{i-1}, a)$$

Identify word meaning m_{d_i} from T
End
For i=1 *to* k
Do For $j = 1$ *to* k
  Do *For each meaning* m_{d_j} *in* d_j
    Do
      Compare meanings m_{d_i} of d_i with other time periods meanings m_{d_j} of d_j
      If Similarity $(m_{d_i}, m_{d_j}) > 0.5$
      Then
      w has same meaning in d_i and d_j and No semantic change occurs between d_i and d_j
      Else
      Semantic change occurs between d_i and d_j *and* w has new meaning in d_i

rows are v context words of word w and column is the time period d_i, and the values are the context words' frequencies in the corpus, the top context words are retrieved from corpus c.

During this step, the word meanings are obtained. Then, the senses are compared to identify the semantic change between the time periods. If new meanings appear through years, then a semantic change has occurred. The result of the semantic change from the proposed algorithm is converted to numerical value, called the (*VSM Semantic Change Algorithm Result*), and it ranges from 0 to 1 according to the meaning comparison results.

The overall value of semantic change is calculated by the averaging the two numerical values as in the following formula:

Semantic Change value= Average (VSM Semantic Change Plotting Result+ VSM Semantic Change Algorithm Result)

Afterward, the collected information from the VSM and the calculated semantic change value are represented as Linked Data using the proposed Arabic Semantic Change Model explained in the next section.

6. Arabic Semantic Change Model

The Arabic Semantic Change (ASC) model is proposed to represent the resulting data of the distributional semantics (VSM) method to identify semantic change as Linked Data. In the ASC model, it was focused to include all the information needed to identify the semantic change. We viewed the model as a vector representation of time periods that includes the distributional semantics information. Each period has a start and end year, a set of top context words, a set of different meanings and the similarity scores between the last period and the referenced period. All the information needed to apply our proposed semantic change identification method and algorithm are represented in ASC model as shown in Figure 7.

The model used existing vocabularies, and new ones were introduced, the new proposed vocabularies and properties are recognized by prefix asc which is the name space or URI used for the model.

In Figure 7, the ontolex:LexicalEntry is connected to blank node asc:SemanticCahngePeriod that represents the semantic change period which is connected to four other nodes. The first node is the time period which is modelled using OWL-Time ("Time Ontology in OWL," 2017), and it has a start and an end time to represent the start and end years of a period of the set of time periods extracted from a corpus (e.g., 2011-2017). In the other cases where the period could be a decade instead of an interval time, the start and end times will have the same value. The second node connected to asc:SemanticChangePeriod is asc:ContextWord that represents the extracted top words in each period, and it is connected to the external dataset (DBpedia ("DBpedia," 2018)) using lemon ontolex:reference property. From the set of the context words, the meanings of a word in each period could be identified. The third node is asc:Meaning that represents the identified meanings from context words in each period, and it is connected to the external dataset (DBpedia) using lemon ontolex:reference property. The VSM algorithm is applied to the meanings to identify the semantic change. The forth node has a value xsd:float an XML Schema Datatype (Carroll and Pan, 2006) with a float value that represents the similarity between the last period vector and the referenced time period vector. These similarity scores can be used to plot the similarity curve to identify the semantic change. Furthermore, in the model the word or the LexicalEntry is connected to an xsd:float that has a range of float values that represents the amount of semantic

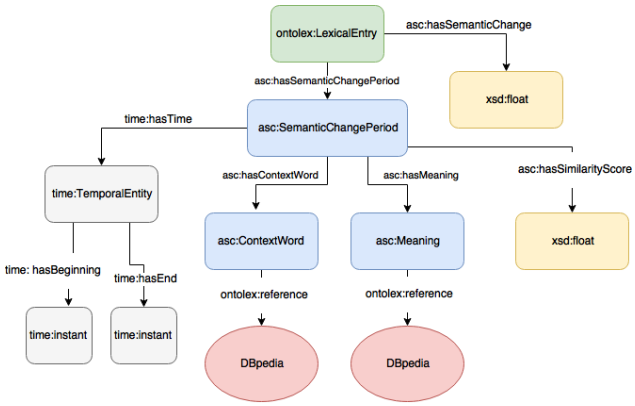

Figure 7 Arabic Semantic Change Model

Table 1: Arabic Semantic Change (ASC) Model

ASC Element	Definition
SemanticChangePeriod	A resource that represents the time period vector and is connected to four other resources
ContextWord	A resource that represents a word's context words retrieved from the corpus
Meaning	A resource that represents the word's different meanings in a Semantic Change Period.
hasSemanticChangePeriod	A property that relates a Lexical Entry with a Semantic Change Period
hasContextWord	A property that relates a Semantic Change Period with the word's Context Word.
hasMeaning	A property that relates a Semantic Change Period with the word's Meaning.
hasSimilarityScore	A property that relates a Semantic Change Period with the Similarity Score float value.
hasSemanticChange	A property that relates a Lexical Entry with the Semantic Change float value.

change occurred. If the value is closer to 1, then a semantic change has likely happened, and if closer to zero, then no semantic change exists, according to meanings and semantic change identification algorithm. Table 1 lists ASC model elements and their definitions.

7. Case Studies

In this section, the Arabic word بئر \be`r\ (well) and the English word (gay) will be used to identify the semantic change occurred to them using VSM algorithm, and to represent their semantic change identification information as Linked Data using Arabic Semantic Change (ASC) Model.

7.1 Arabic Word بئر \be`r\ (well)

The VSM algorithm was used to identify the semantic change of the word بئر \be`r\ (well). First, the 5-gram context words were collected from KACST corpus. Then, the top context words were retrieved, and the similarity scores are computed and plotted using the Arabic Semantic Change tool. Figure 8 shows the plotted similarity curve for the word بئر \be`r\ (well).

From Figure 8, the cosine similarity curve has steep

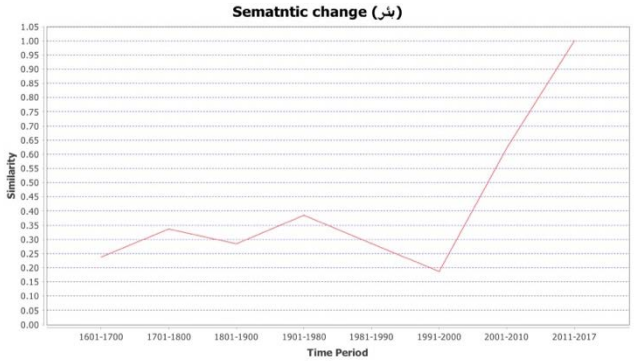

Figure 8 Similarity curve for word بئر \be`r\ (well).

increase and fall in all periods which indicates that the vectors are not similar, and the word has witnessed change through time. Thus, the VSM semantic change plotting result is equal to one. Additionally, the similarity value in the period (2011-2017) is equal to one because we are comparing the period vector with itself.

Table 2 Word بئر \be`r\ (well) top context words and obtained meanings in different time periods

	2011-2017	2001-2010	1991-2000	1981-1990	1901-1980	1801-1900	1701-1800
	غنم	معونة	بضاعة	معونة	معونة	بضاعة	بضاعة
	نفط	رومة	معونة	بضاعة	بضاعة	معونة	مدينة
	مياه	عميقة	رومة	رومة	رومة	رومة	معونة
	ماء	ماء	أريس	زمزم	حفر	زمزم	رومة
	عميقة	عبد	ماء		زمزم		أريس
	حسن	زمزم	جمل				زمزم
	عبد		زمزم				
	عاتر						
	نفطية						
	سلم						
	زمزم						
Identified Meaning	بئر نفط بئر ماء حفرة عميقة	بئر ماء حفرة عميقة	بئر ماء	بئر ماء	بئر ماء	بئر ماء	بئر ماء

Next, the top context words were retrieved using the Arabic Semantic Change tool, and the meanings of the word in different time periods were manually obtained from the list of top context words. Table 2 shows the retrieved top context words and the obtained meanings of the word بئر \be`r\ (well).

Then, by following Algorithm 1, the meanings in all periods were compared and the results were recorded.

Table 3 presents the algorithm comparison results. From observing the senses, the word بئر \be`r\ (well) had as stable meanings of بئر ماء \be`r māa`\ (water well) and حفرة\ḥofrah\ (a dig) from the period (1700-1800) to the period (2011-2017). However, a new meaning arises in the latest period (2011-2017) بئر نفط \be`r nift\ (petroleum well).

Table 3 Word بئر \be`r\ (well) VSM algorithm results

Meaning	2011-2017	2001-2010	1991-2000	1981-1990	1901-1980	1801-1900	1701-1800
بئر ماء \be`r māa`\ (water well)	1	1	1	1	1	1	1
بئر نفط \be`r nift\ (petroleum well)	1	0	0	0	0	0	0
حفرة\ḥofah\ (a dig)	1	1	1	1	1	1	1

This is an indication that the word has witnessed semantic change through time. Thus, the VSM semantic change algorithm result is equal to one. Also, the overall semantic change value is computed by the addition of the two results, the plotting and the algorithm results, and is equal to one.

Furthermore, the Arabic Semantic Change model was used to represent the semantic change identification information and the semantic change value as Linked Data. Figure 9 shows the word بئر \be`r\ (well) Linked Data representation for the time period (2011-2017).

In Figure 9, the similarity score is equal to 1.0 because we are comparing the similarity between the last time period vector with itself. Also, the semantic change calculated using VSM is equal to one. The context words and meanings are presented and linked to the external DBpedia dataset.

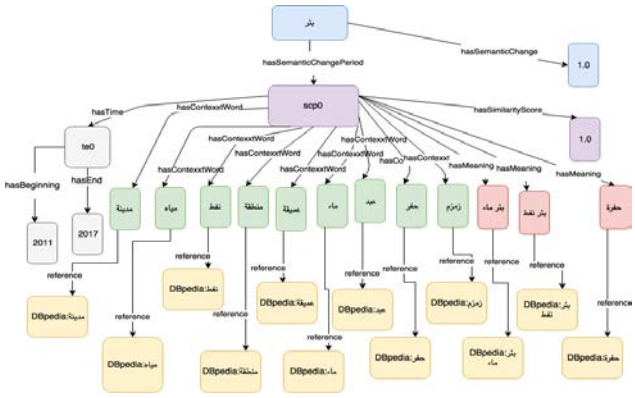

Figure 9 Word بئر \be`r\ (well) Linked Data representation using ASC model

7.2 English Word (gay)

The VSM semantic change method and algorithm were applied to the English word (gay) to identify the semantic change. First, the 5-gram were collected from Google Books corpora. Then, the top context words were retrieved and the similarity curve was plotted using Arabic Semantic Change tool. Figure 10 shows that the similarity curve has a steep increase in the latter decades, which indicates that vectors are not similar and a semantic change has occurred to the word (gay). Thus, the VSM semantic change plotting result is equal to one.

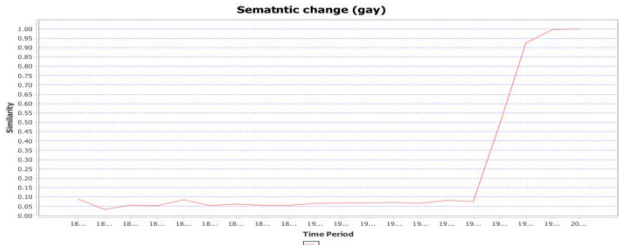

Figure 10 Word (gay) similarity curve

Next, the top context words were retrieved using the Arabic Semantic Change tool and the meanings of the word in different time periods were manually obtained from the list of top context words. Table 4 shows the retrieved top context words and the obtained meanings (last row of Table 3) of the word (gay).

Furthermore, the obtained meanings were compared, and new meanings appeared in the last two decades (lesbian and bisexual) which indicates that a semantic change has occurred to the word (gay). Thus, the VSM semantic change algorithm result is equal to one, and the overall semantic change value is equal to one.

Afterward, the Arabic Semantic Change model was used to represent the semantic change identification information as Linked Data. Figure 11 shows the word (gay) Linked Data representation using ASC model. The similarity score is

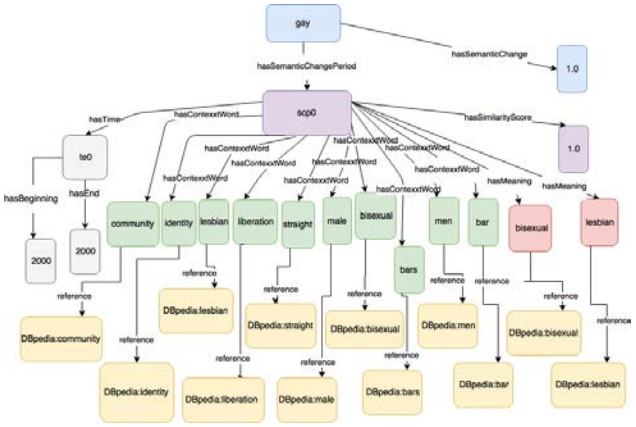

Figure 11 Word (gay) Linked Data representation using ASC model

equal to one because we are comparing the last decade vector with itself in other decades this value varies.

8. Conclusion

In this paper, the modified Vector Space Model (VSM) algorithm to identify semantic change in the Arabic language was presented. Also, The Arabic Semantic Change (ASC) Model to represent as Linked Data the semantic change identification information using VSM was proposed. The algorithm and model were evaluated using Arabic and English words. Therefore, the proposed ASC model could be used with any language as shown in this paper. The words represented using ASC could be expanded and published as a Linked Dataset.

9. Acknowledgments

This project is granted by King Abdulaziz City for Science and Technology (The Graduate Studies Grant Number 37-1233). We appreciate Dr. Abdulmohsen Al-Thubaity exceptional support and insights.

Table 4 Word (gay) top context words and manually obtained meanings in different decades

2000	1990	1980	1970	1960	1950	1940	1930	1920	1910	1900	1890	1880	1870	1860	1850	1840	1830	1820	1810
liberation	liberation	liberation	lively	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant
male	male	movement	happy	noisy	blades	nineties	charming	flowers	flowers	flowers	plumage	plumage	fantastic	fantastic	fantastic	fantastic	fantastic	rose's	rose's
bars	bars	community	flowers	blades	spirited	charming	flowers	throng	throng	animated	flowers	flowers	flowers	flowers	flowers	flowers	flowers	dissolute	dissolute
identity	couples	male	colors	nineties	nineties	flowers	laughter	laughter	laughter	throng	throng	throng	throng	animated	animated	animated	animated	fantastic	sprightly
couples	identity	bars	men	charming	charming	hearted	hearted	careless	apparel	laughter	apparel	careless	careless	throng	sprightly	sprightly	sprightly	flowers	witty
bar	bar	rights		flowers	flowers	laughter	careless	laugh	careless	hearted	careless	laugh	laugh	sprightly	laugh	laugh	careless	animated	dissipated
straight	straight	identity		laughter	laughter	colorful	laugh	colors	laugh	apparel	laugh	joyous	joyous	laugh	joyous	joyous	dissipated	sprightly	men
men	men	bar		colorful	hearted	witty	colors	men	joyous	careless	joyous	colors	colors	careless	dissipated	dissipated	lively	careless	lively
lesbian	lesbian	straight		witty	colorful	laugh	men	lively	colors	laugh	colors	lively	graceful	joyous	colors	colors	happy	vermeil	thoughtless
movement	movement	males		laugh	witty	men	lively	happy	men	joyous	lively	happy	lively	colors	graceful	lively	gaudy	dissipated	sportive
community	community	men		men	laugh	colors	happy	cheerful	lively	colors	happy	thoughtless	happy	graceful	lively	happy	thoughtless	lively	cockade
bisexual	bisexual			colors	colors	lively	cheerful	colored	happy	lively	thoughtless	cheerful	thoughtless	lively	happy	thoughtless	glittering	happy	dresses
rights	rights			lively	men	friendly	colored	coloured	cheerful	happy	cheerful	fashionable	cheerful	happy	thoughtless	glittering	cheerful	thoughtless	attire
				friendly	friendly	happy	coloured	festive	colored	cheerful	fashionable	colorrd	fashionable	thoughtless	glittering	cheerful	fashionable	glittering	courtier
				happy	lively	cheerful	festive	ribbons	coloured	colored	colored	gorgeous	colored	cheerful	cheerful	fashionable	attire	cheerful	giddy
				cheerful	happy	colored	flags	flags	festive	festive	festive	festive	gorgeous	colored	fashionable	gorgeous	brilliant	fashionable	licentious
				colored	cheerful	festive	brilliant	brilliant	ribbons	flags	flags	flags	festive	fashionable	gorgeous	dresses	colours	attire	motes
				festive	colored	carefree	colours	colours	flags	attire	dresses	dresses	flags	gorgeous	dresses	attire	licentious	amusing	fantastic
				carefree	carefree	flags			attire	brilliant	attire	attire	dresses	dresses	attire	brilliant		colours	flowers
				flags	festive	brilliant			brilliant		brilliant	brilliant	attire	attire	brilliant	colours		licentious	enamell
				brilliant	flags	amusing			colours				brilliant	brilliant					animated
				amusing	amusing								merry						careless
				colours	brilliant														vermeil
					lighthearted														joyous
																			happy
																			glittering
																			cheerful
																			fashionable
																			amusing
																			colours
Lesbian bisexual	Lesbian bisexual		Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful

REFERENCES

Abele, A., McCrae, J. P., Buitelaar, P., Jentzsch, A., & Cyganiak, R. (2017). Linking Open Data Cloud Diagram. [cited at: 20-9-2017}, Retrieved from http://lod-cloud.net/

Abuhadeemah, T. (2008). *Studies in Arabic Dictionaries and Semantics* Riyadh, Saudi Arabia: Dar Almarefah for Human Development

Al-Thubaity, A. O. (2015). A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction. *Language Resources and Evaluation*, *49*(3), 721-751.

Beckett, D. (2004). RDF/XML Syntax Specification [cited at: September 20, 2017}, Retrieved from http://www.w3.org/TR/REC-rdf-syntax/

Bizer, C. (2009). The Emerging Web of Linked Data. *Intelligent Systems, IEEE*, *24*(5), 87-92. doi:10.1109/mis.2009.102

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*. doi:10.4018/jswis.2009081901

Carroll, J. J., & Pan, J. Z. (2006). XML Schema Datatypes in RDF and OWL. [cited at: March 1, 2018},

Chiarcos, C., McCrae, J., Cimiano, P., & Fellbaum, C. (2013). Towards Open Data for Linguistics: Linguistic Linked Data *New Trends of Research in Ontologies and Lexical Resources* (pp. 7-25): Springer.

Chowdhury, G. G. (2010). *Introduction to Modern Information Retrieval*: Facet publishing.

DBpedia. (2018). [cited at: March 1, 2018},

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, *41*(6), 391.

Fellbaum, C. (1998). *WordNet*: Wiley Online Library.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. *Studies in linguistic analysis*.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Human Factors and Behavioral Science: Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems. *The Bell System Technical Journal*, *62*(6), 1753-1806.

Golub, G. H., & Reinsch, C. (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische mathematik*, *14*(5), 403-420.

Gulordava, K., & Baroni, M. (2011). *A Distributional Similarity Approach to The Detection of Semantic Change in the Google Books Ngram Corpus.* Paper presented at the Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics.

Harris, Z. S. (1954). Distributional Structure. *Word*, *10*(2-3), 146-162.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1* (1st edition ed.): Morgan & Claypool.

Huang, A. (2008). *Similarity Measures for Text Document Clustering*. Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

Hurford, J. R. (2007). *Semantics: A Coursebook*: Cambridge University Press.

Issa, F., & Issa, R. (2008). *Semantics the Theory and the Application*. Alexandria, Egypt: Dar Almarefah Algameiah.

Jatowt, A., & Duh, K. (2014). *A Framework for Analyzing Semantic Change of Words Across Time*. Paper presented at the Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries.

Khan, F., Boschetti, F., & Frontini, F. (2014). *Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources*. Paper presented at the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing.

Löbner, S. (2002). *Understanding Semantics*: Taylor and Francis Group.

Lyons, J. (1977). *Semantics* (Vol. 53): Cambridge University Press.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Río, J. G. d., Hollink, L., Montiel-Ponsoda, E., & Spohrx, D. (2012a). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, *46*(4), 701-719.

McCrae, J., Montiel-Ponsoda, E., & Cimiano, P. (2012b). Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*, 25-34.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., & Orwant, J. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *science*, *331*(6014), 176-182.

Qalalah, A. (2017). *Semantic Development Signs and Issues: A Study in Language Measures for Ibn Fares*. Irbid, Jordan: Alam Alkutob AlHadeeth.

Rodda, M. A., Senaldi, M. S., & Lenci, A. (2016). Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *CLiC it*, 258.

Saeed, J. I. (1997). Semantics: Oxford: Blackwell Publishing.

Time Ontology in OWL. (2017). [cited at: November 7, 2017}, Retrieved from https://www.w3.org/TR/owl-time/ -time:TemporalEntity

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, *37*, 141-188.

van Aggelen, A., Hollink, L., & van Ossenbruggen, J.

(2016). *Combining Distributional Semantics and Structured Data to Study Lexical Change.* Paper presented at the European Knowledge Acquisition Workshop.

Weaver, W. (1955). Translation. *Machine translation of languages, 14*, 15-23.

Wijaya, D. T., & Yeniterzi, R. (2011). *Understanding Semantic Change of Words Over Centuries.* Paper presented at the Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web.