University Politehnica of Bucharest

Faculty of Automatic Control and Computers

Computer Science Department

Discourse Analysis based on Semantic Modelling and Textual Complexity

- Summary of the Ph.D. Thesis -

Scientific Advisor

Ştefan Trăuşan-Matu, Ph.D.

Ph.D. Candidate

Marius-Gabriel Guțu

Bucharest

2017

Table of Contents

1 Introduction

Many documents are being created every second in the world, of which online dissemination produces nowadays 2.5 Exabytes of data[1] (2.5 billion Gigabytes). The 2013 IBM Annual Report states that 80% of the data produced every day represents "unstructured" data[2], which consists of images, videos, audio materials, social media posts, data collected from Internet of Things devices or other kinds of data. The unstructured data brings new areas of exploration in that the analysis of such input can lead to a better understanding of human needs by extracting knowledge and producing results that may lead to a world where societies' problems diminish and people focus on their essential needs and interests (John Walker, 2014). Of the data produced every date just a small part consists of text. Of them, we mention: more than 200 million e-mail messages sent per minute[3], almost 2.5 million shares on Facebook and more than 277,000 tweets sent on Twitter[4].

Text represents valuable data that can lead to interpretations that might be hardly noticed by the human eye or could be barely determined by the humans' cognitive processes by a first reading. In short, what a human could achieve through an in-depth analysis of a document could be automated through mechanisms of textual analysis. These mechanisms have been intensively studied through the last decades (Manning & Schütze, 1999) and they consist the Natural Language Processing (NLP) field in computer science. Natural Language Understanding (NLU) is a NLP field that rely on interpretation of texts through reading comprehension. A notable researcher in this field was Terry Winograd, who is one of the first researches to develop a computer system that could interpret texts (written in English) so that to answer questions and perform reasoning (Winograd, 1972). Winograd's experiments were brought into practical application by his SHRDLU computer program that allowed the users to perform the movement of a robotic arm through textual commands (Winograd, 1980).

[1] http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/
[2] https://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf
[3] https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/
[4] https://www.domo.com/blog/data-never-sleeps-4-0/

1.1 Goals of the thesis

This thesis goal is to investigate the process of analysis of textual contents by using different types of documents. Neither of the sources mentioned above (e-mail, Facebook, Twitter) were used as they usually contain shallow information with regards either to the length of the text or to the quality of the content. Regarding the length of the content, commonly users share small pieces of text on these platforms. With respect to the quality, much of the texts shared through these platforms do not bring so much information. We targeted more elaborated texts like scientific papers, students' assignments or CVs. There is a high need for analysis and extraction of valuable information from documents like these ones in that to automate or ease processes. Scientific papers could be automatically classified into their most appropriate category or the keywords might be extracted to help the researchers determine whether they covered the requirements of a conference's or journal's topic. Students' assignments could be automatically assessed to allow the professor to focus on more essential information such as students' involvement, restructuring materials by mass customization to fit groups of students based on their knowledge. An applicant for a job opening may produce a better CV focused on the employer's needs by adapting the content based on data extracted by a dedicated tool. In contrast, an employer could automatically determine who is the best candidate for their position, thus saving money by focusing on interviewing the most suitable applicants.

The thesis is focused on the development of innovative methods able to analyze documents through automated NLP techniques that rely on semantic models and textual complexity to automate the laborious work performed by humans. By integrating multiple of the services together, advanced information could be extracted and more processes could be automated to remove the tiresome work that destroy individuals' capacity of creation.

1.2 Thesis Novelty

The novelty of this thesis is framed by the integration of means of discourse analysis into everyday scenarios that require people to perform repetitive work. The automation of processes that involve extraction of information from documents (like scientific papers, CVs or chat conversations) would ease the work by providing higher level data that could be used to extract additional information. The experiments were aimed at discovering documents' particularities and characteristics through the integration of NLP techniques for performing discourse analysis based on semantic models and textual complexity.

Part of the experiments includes the analysis of chat conversations to determine the best mechanisms to detect implicit links. Implicit links are used in advanced processes like detection of topics, assessment of knowledge building and of participant integration. The analysis of chat conversations is in its birth years and allows for novel experiments to be conducted. The investigation of the most important characteristics that surmise whether a CV is suitable for a position in terms of visual aspect or textual content is another idea that has been scarcely studied recently. Such a service could provide benefits for companies looking for hiring "the ideal candidate". The development and validation of a semantic annotation tool that automatically classifies scientific papers or other types of documents is of great use for many companies. Although needing an initial "setup" to provide a list of categories, the automated work that follows allows the ease of work by completely removing the process of classification and allowing companies to focus on other tasks that are performed after categorization is performed.

The development of the online version of *ReaderBench* and the integration of demo clients for the services integrated on the server allows simplified usage for regular users. Moreover, the exposure of the *ReaderBench* API grants access for developers to create their own applications that rely on data provided by the *ReaderBench* framework. The maturity of the services come in that much of them were validated through experiments. Although most experiments were performed on English, the transition to other languages could be performed with ease.

1.3 Thesis Structure

This thesis is structured into two parts. Part I covers theoretical aspects and state of the art related to the NLP field. This include the description of the *ReaderBench* framework and the most recent studies related to discourse analysis and Computer Supported Collaborative Learning (CSCL). Part II presents the empirical experiments that were conducted, structured in two sections: 1) experiments relying on discourse analysis in general; 2) experiments relying on detection of implicit links in CSCL chat conversations. Next follows a brief presentation of the services involved in the experiments, which were exposed through an Application Programming Interface. The thesis continues with discussions together with advantages of the described experiments and the identified problems, followed by conclusions outlining personal contributions of the thesis and directions for future research.

Part I. Theoretical Framing

2 Computational Discourse Analysis

2.1 Overview

This thesis is aimed at performing discourse analysis of written texts to extract valuable information that people can rely on to automate processes or to help them in decision making. A scenario covers the automated assessment of students' activity throughout courses in a Massive Open Online Course (MOOC) platforms. By automated interpreting of their assignments the students may be scored in an objective manner, suitable for diminishing teachers' subjectivity. The learning materials might be adapted to the students' ability to learn new concepts through mass-customization (Nistor, Dehne, & Drews, 2010) and fit their needs. Software applications might automatically categorize collections of documents into their most relevant categories by applying machine learning algorithms (Sebastiani, 2002) or classical clustering methods (Hanson & Bauer, 1989) through discourse analysis.

Texts from CSCL chat conversations outline another scenario that involve discourse analysis for determining the most relevant keywords, extracting the topics of discussion or interpreting participants' interactions to determine who brought the most knowledge or who was the most communicative. Such information might be hardly detected by human experts and might be prone to subjectivity. CV documents allow people to describe themselves for a job position. Automated analysis of the CV may provide recommendations for improvement to increase their chances for the job. For the company itself, an automated analysis of CVs provided by applicants may ease the process of selection by providing the most suitable people to be interviewed, thus reducing the costs and time for examining the entire lot of candidates.

While these are just some examples, there are many types of documents that could be automatically analyzed; the involved techniques rely on NLP through semantic models to restraint data into meaningful information. The process would usually still involve the involvement of human thinking and reasoning, but this would occur at a higher level, that discharges subjectivity and allows humans to center on the procedures that require implication of experts of a specific field of interest. Automated processes should provide more details for the materials that are scarcely understood by individuals; the materials should be taught on a broader scale where individuals do need to know more information regarding a specific domain.

2.2 Natural Language Processing

Natural Language Processing (NLP) techniques (Manning & Schütze, 1999) are more and more used nowadays since they provide accurate and efficient analyses of written texts. The particularities of the documents allowed the envisioning of the stages required for a "cleaner" text, which represents a text that contains only meaningful words regarding semantics, thus a more relevant text to be interpreted by machines. Lemmatization was identified as a necessary process because of the large number of forms that words may have – singular, plural, articulated, un-articulated, etc.

Most of the processes that use NLP rely on text cohesion, property of a text to be held together through grammatical and lexical linking and to have a meaning. Text cohesion can be computed with the help of ontologies and semantic models. Ontologies use dictionaries and relations between words, while semantic models count on a collection of documents that is used for training and leads to the creation of a model that allows the computation of a "similarity score" between two units of text.

2.3 Discourse Analysis

The Cambridge English Dictionary defines *discourse* as "spoken or written discussion"[5]. The idea of examining textual content in relation to spoken dialogue was brought to the attention of the scientific community by Bakhtin since the '80s (Bakhtin, 1981). Bakhtin introduced the concept of *voices* and he interpreted how these voices blend to form a cohesive discourse. He studied the concept of discourse be it written text or spoken dialogue to discover resemblances between the two on one hand and to explore particularities of each one on the other hand. His compelling observations was that written text mirror spoken dialogue by showing most of the characteristics of the latest one regardless the type of the text: novel, story, narrative or written dialogue. Discourse structure consists of connectives and metrics derived from polyphonic model of discourse (Dascalu, 2014), which considers the evolution of points of view and provide insights in terms of the text's degree of elaboration. Word features and vectors from the integrated linguistic resources are also used to reflect specific discourse traits.

[5] http://dictionary.cambridge.org/dictionary/english/discourse

2.4 Semantic Models

When it comes to determining the semantic similarity between two words, then the following question arises: *how can we evaluate it?* Techniques promoted through researches within the latest years usually quantify the similarity with scores that express either strong semantic relations for higher values or delicate or nonexistent semantic relations between words for smaller values. Extrapolating the interpretation, we may notice that in some cases words denoting high similarity scores might be either synonymous or words within the same lexical field. Words denoting low similarity scores might express either concepts that have no relevance one to another or words that are "far" one from each other in their lexical chains.

The differentiation among the two types of semantic relations may be hard to be made, particularly when relying on pre-trained semantic models based on distributional semantics, as we will further see. These approaches depend on the training corpus and determine similarity scores with regards to the probability of the words to occur together, which of course is much higher in this case than in pure linguistic approaches that usually rely on lexicons.

2.4.1 Ontologies and Distributional Semantics

A lexicon is referred in linguistic sciences as the vocabulary of a person, a language in general or a branch of a knowledge base. Together with a grammar, which defines a system of rules that allow the combination of words to create meaningful sentences, they define a spoken language. Considering that each language has its rules for generation of words and relations among concepts, the necessity of a global interest to support this necessity appears. As English is one of the most spoken languages in the world, there are many resources and software tools related to semantics available. Most of the experiments were performed for English data.

Distributional semantics studies semantic similarities between linguistic elements based on their distribution within a collection of documents, which usually contains many documents and is called corpus of documents. Distributional semantics is constructed on the distributional hypothesis idea, which considers that words occurring together in the same contexts have similar meaning (Harris, 1954). This idea was further promoted by the English linguist Firth in that a word is defined "by the company it keeps" (Firth, 1957). The more similar two words are in terms of semantic meaning, the more chances for their occurrences to appear in similar contexts (Yarlett & Ramscar, 2008).

2.4.2 WordNet

WordNet (Miller, 1995) is a large lexical database that was initially developed for English at Princeton. It can be used through the online version or as standalone software application. WordNet's database files can be used by programmers to extract information and relations between words to perform linguistic analysis and to fulfill NLP-related researches. In its internal representation, words are organized into so-called *synsets*, which are sets of similar words. Relations are mapped within these words in a hierarchical representation of a tree that express hypernym-hyponym relationships. Based on its internal representation, several distance algorithms were developed.

For other languages, researchers have developed dedicated WordNet versions. Thus, for French there are three versions available[6], of which we chose WOLF (Wordnet Libre du Français, Free French Wordnet) (Sagot, 2008) for our experiments that involved the French language because it is distributed within a package containing a large number of WordNet dictionaries provided multiple languages (Bond & Foster, 2013).

2.4.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) is one of the most mature and still most utilized statistical semantic model. LSA is used for representing relationships between the concepts covered by a collection of documents and the terms contained within these documents. The established relations refer to semantic similarity, as LSA was developed over the distributional hypothesis.

By using a training corpus, LSA extracts a set of concepts and generates a term-document matrix containing the number of occurrences of each word per paragraph or per document (the level of granularity depends on the characteristics of the experiment), unique words being represented on the matrix' rows, while paragraphs being stored on its columns. A singular-value decomposition (SVD) (Golub & Reinsch, 1970) is further performed to reduce the number of words at the same time maintaining the columns, followed by a reduction of the matrices' dimensionality through a projection on k dimensions in order to determine indirect links induced between groups of terms and underlying documents. Based on a vector space model that highlights co-occurrences of words within documents, the similarity score between two words is computed as the cosine of the angle formed by the two corresponding rows. LSA uses a "bag of words" approach that disregards words' order.

[6] http://globalwordnet.org/wordnets-in-the-world/

2.4.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a generative probabilistic process built on top of the assumption that documents integrate multiple topics and can be therefore considered a mixture of corpus-wide topics. Each topic represents a Dirichlet distribution (Kotz, Balakrishnan, & Johnson, 2000) over the vocabulary where related concepts have similar probabilities based on co-occurrence patterns from the training corpora. Although each topic contains all the words from the vocabulary, a clear differentiation in terms of corresponding probabilities can be observed between salient versus dominant concepts. Similar to LSA, LDA relies on the "bag of words" approach and classifies new texts in terms of the latent topics inferred from the model trained on a text collection. Documents and words alike become topics distributions drawn from Dirichlet distributions, while semantic similarities between textual fragments are determined using the Jensen-Shannon dissimilarity (JSH) (Manning & Schütze, 1999), a symmetric smoothed alternative of the KL divergence (Kullback & Leibler, 1951).

2.4.5 Word2vec

Word2vec is a one of the newest NLP semantic models used for computing text cohesion between documents. It is a technique developed for assessing the semantics of a text, proved in recent years to provide better performance for several NLP tasks involving semantic analysis than previous approaches (Mikolov, Chen, Corrado, & Dean, 2013; Swoboda, Hemmje, Dascalu, & Trausan-Matu, 2016). Word2vec is one of the most recent methods used for representing words and phrases in a vector-space model within a limited number of dimensions, called word embeddings, which are computed using a neural network model. The resulted embedded space can be used afterwards to compute a semantic similarity between words and phrases. In the case of Word2vec, each embedding is computed using the context before and after each word occurrence in the training dataset. This way, words co-occurring in similar contexts are represented closer in the embedded space, while words that do not share similar context are represented in different regions of this space (are farther apart). These embeddings are computed using a neural network model which can process larger volumes of text than any of the previous methods that compute word embeddings or vector representations. The resulted embedded space can be afterwards used to compute a semantic similarity between words and phrases.

2.4.6 Text Cohesion and Cohesion Network Analysis

The idea of quantifying the semantic similarity between two textual units has been extensively studied in the NLP field within the last years. Semantic cohesion reflects the degree to which two text fragments are related one to another in terms of meaning (Bestgen, 2012) and can be automatically evaluated using several approaches. In previous studies in the NLP field, several techniques gained high popularity. The first one consists of applying different semantic distance functions on ontologies (Budanitsky & Hirst, 2006), such as the WordNet lexical database (Miller, 1995). Second, Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) is the most frequently used method to compute semantic similarity by relying on vector spaces of keywords (terms). Third, probabilistic topic modeling has gained an increasing attention lately, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) being the most frequently used method of this kind.

Cohesion Network Analysis (CAN) introduced a generalized model based on the cohesion graph to represent discourse structure and underlying cohesive links. Based on CNA, a topic mining module was implemented, which extracts the most relevant concepts from a text. Integrated within the web interface, this module draws a concept map of these keywords: the nodes represent the central topics and the links between them depict the semantic similarity between two concepts; the size of each node is proportional to its relevance.

2.5 The ReaderBench Framework

The *ReaderBench* framework (Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014; Dascalu et al., 2015a, 2015b; Dascalu et al., 2015c), comprises of advanced NLP techniques based on Stanford CoreNLP (Manning et al., 2014) used to expose a wide variety of language services. We can consider our framework as being unique as it provides a unitary core engine centered on cohesion and on dialogism (Dascalu, Trausan-Matu, Dessus, & McNamara, 2015a, 2015b), the latter being reflected in the implemented polyphonic model (Trausan-Matu, Stahl, & Sarmiento, 2007). Multiple connected services addressing different facets of comprehension assessment and prediction are thus deployed. Tutors are capable to perform an apriori assessment of learning materials, but also to evaluate a posteriori learner's written traces consisting of essays, self-explanations or utterances in CSCL conversations. All these services are described in detail in subsequent sections.

ReaderBench uses documents and meta-documents to store texts in its internal representation (Dascalu et al., 2015a) that considers a multi-hierarchical graph representation including paragraphs, sentences and words. A meta-document is a document that contains sections with headings and content, in which content represents a document. Relations between different hierarchical levels, for example between one sentence and the entire document, are also mapped within our cohesion graph. This process is called Cohesion Network Analysis (CNA) (Dascalu, McNamara, Trausan-Matu, & Allen, 2017; Dascalu, Trausan-Matu, McNamara, & Dessus, 2015) and supports the majority of *ReaderBench* services (Dascalu et al., 2015a). The CNA is performed by using semantic models. In *ReaderBench's* internal representation, the previously mentioned meta-document is a distinct type of document that contains multiple sections, with each section having a title and multiple paragraphs. Each section may contain subsections on a multi-level approach where the most upper sentence represents the meta-document itself, while the most lower sections represent general documents.

For the transformations of text into meaningful data for machines, *ReaderBench* uses a NLP pre-processing pipeline (Dascalu, 2014) that consists of tokenization, text splitting, part of speech tagging, lemmatization, named entity recognition, dependency parsing and co-reference resolution. All these steps enclosed within our pre-processing pipeline are helpful in sanitizing the input text that will be further used both for training semantic models and for performing natural language processes.

2.6 Keyword Mining

Keyword Extraction has applicability in various NLP scenarios (Lahiri, Choudhury, & Caragea, 2014) of which the detection of primary topics, also called Topic Mining, is one of the most frequent. Through Topic Mining researchers or regular users can determine whether they need to use fewer occurrences of specific words or whether they need to increase the number of occurrences for others which are more important, but are shadowed by common words. An opposite process is the generation of keywords of a text.

"Classic" methods of determining keywords of a text make use of its number of occurrences, which is usually computed through the *tf-idf* (term frequency-inverse document frequency) (Dascalu et al., 2015b). However, the keywords with the highest numbers of occurrences do not always map the most relevant keywords from a text. Thus, a mechanism to remove irrelevant words showed as necessary to be integrated. Irrelevant words were considered as being the most common words, which brought the idea to determine a "commonness" score for each word. (Savický & Hlavácová, 2002). These techniques are based on the fact that relevant words can be found uniformly distributed within a text. The most stable formula is called Average Logarithmic Distance (ALD) and determines the relevance of a word as shown in (1), where N represents the total number of words of the text, f represents the number of occurrences of the word in question and d_i represents the *distance* of the i-th occurrence of that word.

$$ALD = \frac{1}{N}\sum_{i=1}^{f} d_i \log_{10} d_i \quad (1)$$

Distances are computed as shown in (2) for every i between 2 and f in which n_i represents the position of the i-th occurrence of the word in the text.

$$d_i = n_i - n_{i-1} \quad (2)$$

The distance of the first occurrence of a word is computed as shown in equation (3), which computes the distance between the beginning of the text and the current position summed with the distance from the last occurrence to the end of the text.

$$d_1 = n_1 + (N - n_1) \quad (3)$$

Thus, the distance of a word with only one occurrence is the entire length of the text. The positions of words are calculated as the index of the word in a set containing all the words from the text in their occurrence order.

2.7 Textual Complexity

Textual complexity allows the extraction of valuable information that might provide relevant insights regarding a text. These can include their ability to understand a text, their capacity to elaborate complex texts, or their competence within a specific domain. Complexity indices include statistic surface indicator (e.g., average paragraph, sentence or word lengths, number of commas, word and character entropy), syntax factors (statistics on different parts of speech, average number of first, second or third person pronouns per paragraph, depth of parsing tree) and semantic cohesion. Assessing textual complexity is a difficult, but important endeavor, especially for adapting learning materials to students' and readers' levels. With the continuous growth of information technologies spanning through various research fields, automated assessment tools have become reliable solutions to automatically assessing textual complexity. *ReaderBench* integrates a multitude of indices ranging from classic readability formulas, surface indices, morphology and syntax, as well as semantics and discourse structure.

Surface indices are the simplest measures that consider only the form of the text. This category includes indices such as sentence length, word length, the number of unique words used, and word entropy. These indices rely on the assumption that more complex texts contain more information and, inherently, more diverse concepts. Word complexity indices focus on the complexity of words, but goes way beyond their form. Thus, the complexity of a word is estimated by the number of syllables and how different the flectional form is from its lemma or stem, considering that adding suffixes and prefixes increases the difficulty of using a given word. Moreover, a word's complexity is measured by considering the number of potential meanings derived from the word's senses available in WordNet, as well as a word's specificity reflected in its depth within the lexicalized ontology. Syntactic and morphologic features are computed at the sentence level. The words' corresponding parts of speech and the types of dependencies that appear in each sentence can be used as relevant measures, reflective of a text's complexity. In addition, named entity-based features are tightly correlated with the amount of cognitive resources required to conceptualize the given text. Semantic cohesion plays an important role in text comprehension and our framework makes extensive usage of Cohesion Network Analysis. *ReaderBench* estimates both local and global cohesion by considering lexical chains, different semantic models (semantic distances in WordNet, LSA, LDA , and Word2vec), as well as co-reference chains.

3 Implicit Links in CSCL conversations

3.1 Overview

Computer Supported Collaborative Learning (CSCL) is a process in that participants discuss with regards to a specific topic through a software application. The goal of the communication is to share knowledge on one hand and to solve a problem in a collaborative manner on the other hand. The communication is aimed at solving problems hard to be solved by individual efforts or at reaching an agreement. CSCL is implemented through software applications that might be either synchronous or asynchronous. Synchronous applications refer to software were participants receive only the contributions that are added when they are connected, while asynchronous applications allow the retrieval of contributions that were added even when the participants were disconnected. Examples of synchronous applications are online chats and instant messaging platforms. Examples of asynchronous application are forums and blogs.

3.2 Computer Supported Collaborative Learning

The advance of communication and collaborative technologies on the social web (Stahl, 2006) lead to an increased concern with regards to Computer-Supported Collaborative Learning (CSCL). CSCL emerged as a well-suited method for learning through a knowledge building process according to the socio-cultural paradigm (Stahl, 2006). One of the most popular technologies used in CSCL is instant messenger (chat). Chat environments, when also integrating explicit referencing facilities, enable small groups of students to generate complex parallel threads of discussions, inter-animating in a polyphonic framework (Trausan-Matu & Rebedea, 2010).

Stahl conducted multiple case studies in on collaborative technologies and analysis of interactions through the Virtual Math Teams project, were students solved mathematical problems in a collaborative online platform (Stahl, 2006). Thus, he showed that even topics that rely on more advanced knowledge or cognitive processes like demonstrations and computations processes can be more easily solved together by involving collaborative technology; this lead to more technologies to adapt the CSCL paradigm.

3.3 The Polyphonic Model of Discourse

Polyphony represents a model for human communication that is performed either in natural language using words or through nonverbal communication using gestures and has applicability in small teams through collaborative technologies (Trausan-Matu, Stahl, & Zemel, 2005). Polyphony is a concept that initially appeared in music and later has been transferred into computer supported tools where multiple participants interact to solve a problem, but at the same time each one maintains their individuality.

The polyphonic weaving of knowledge construction in CSCL chat conversations involves threads composed of *explicit* and *implicit links*. The latter are pairs of utterances part of a discussion thread logically connected in a discursive structure. Implicit links may be detected using NLP techniques: repetitions, lexical chains, adjacency pairs of speech acts (Trausan-Matu & Rebedea, 2010), and semantic models or other means for measuring the similarity between two utterances (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015).

3.4 Dialogism

Dialogism refers to the multiplicity of points of view yielding as the basis for framing a comprehensive model of discourse (Bakhtin, 1981). Dialogism is present both in written texts and in oral speaking. Bakhtin showed that everything that is being said is influenced by previous sayings, it also influences further dialogue, but, the more interesting part, what one individual is saying is influenced by the possible responses that could occur, too. Moreover, every speech act of a discourse can be considered a dialogue, even a single utterance. Every speech act does also generate another utterance and it represents by itself an entity that contains enough information (comprised of ideas and ideologies) to be independently interpreted in that why that all the other utterances satisfy the same property. Thus, the utterance is considered the basic unit of a conversation.

Bakhtin claims that a participant of a conversation gives up his *turn* to let another participant to contribute to the conversation, while the latter can bring new ideas, can generate new topics, or may respond the contribution of the first participant maintaining the current topic of discussion. When we say that a participant *replies* to another participant's contribution we think to any form of interaction, such as confirmation, denial, correction, empathy, etc.

3.5 Explicit and Implicit Links

Implicit links detection in chat conversations represents an important research topic due to the lack of referencing options availability in the clear majority of online chats, as well as discourse segmentation that makes the discourse difficult to follow. In contrast to other types of web interaction tools like forums or social media, chats do not provide any "*reply-to*" option. Thus, they do not provide any flow in the discussion; therefore, manual insertions or automated annotations are required to conduct advanced analyses such as topic detection, lexical chains extraction or evaluating the degree of collaboration in problem solving task.

Implicit links established between contributions of a conversation outlines the concept of references. They may not occur immediately, but they can appear at some chronological distance by the original contribution. However, it has been showed that most of the times references are established in the proximity of the referenced contribution. Modern techniques of linguistic analysis of text can be used for this purpose. A fundamental step in detection of implicit links between contributions of a chat conversation is represented by the semantic similarity between the words which the contributions are composed by Gutu, Rebedea, and Trausan-Matu (2015) shows a comparison of similarity scores between pairs of words extracted from a corpus (a collection of written texts) of about 200 chat conversations. The paper focuses on a comparison of similarity scores obtained using methods of computation of semantic similarity, namely text-based methods, which requires previous training on a specific set of data and WordNet-based techniques, which use WordNet, the English lexical database (Miller, 1995). The paper shows that the choice of training corpus according to the text is crucial for achieving good results for corpus-based techniques and that for the data used in the experiment the scores obtained using WordNet-based methods generated more natural results, being very correlated with manual annotations of human experts.

The detection of implicit links in chat conversations consists an essential part of this thesis, thus several research experiments were performed and are further described in this chapter. This research topic has application in many scenarios, of which we mention analysis of knowledge building, assessment of participant interaction, detection of the most important topics involved in the conversations, evaluation of voice overlaps and many others. The implicit links may also sow how one participant's ideas are affected by other participants' ideas, how their ideas bind together and how they are transmitted from one participant to another.

Part II. Empirical Studies

4 Practical Applications of Automated Discourse Analysis

4.1 Overview

This section presents five case studies with regards to automated discourse analysis based on semantic models and textual complexity. The first two case studies refer to automated classification of scientific papers relying on textual content on one hand and on keywords on the other hand. The experiments were conducted using LSA and LDA and were aimed at demonstrating that the tedious work of categorization can be performed by machines with the scope of allowing people to focus on more important tasks. Next follows the presentation of a case study regarding a tool build for automated assessment of the quality of a CV. While individuals cannot determine with ease whether their CVs are "good" or "bad", this tool learnt by using a training corpora what characterizes the quality of a CV. The chapter continues with a description of two educational scenarios: one referring to the integration of NLP facilities into a MOOC platform. The goal was to provide keywords and allow students to search through course materials. The following scenario was aimed at better depicting categories to be used for a collection of learning materials. Assessment of students' interventions within a Massive Open Online Courses (MOOC) platform could be another practical example. This was studied in an experiment aimed at personalization of the content of a medical MOOC platform so that to fit group of users (Nistor et al., in press). Another application on the MOOC platforms was presented in an experiment aimed at showing practical examples of integration within an entire course from Grenoble University (Dessus, Gutu, Dascalu, Diouf, & Trausan-Matu, 2017). Here, using the *ReaderBench* framework the most important concepts of the learning material could be extracted and the students were provided the keywords ordered by importance together with a visual representation through a Concept Map.

According to each experiment, different input data were used, namely the kind of file and the structure of the inner text together with the inner representation in the software applications. Both raw text files and advanced data such as PDF files were used. In the experiments, the PDF files consisted of CV files and scientific papers.

4.2 Case Study 1: Text Categorization using Cohesion Network Analysis

A semantic annotation tool was developed to provide recommendations regarding categories that should be used for automated labelling. To this aim, the SemEval-2010 task 5 (Kim, Medelyan, Kan, & Baldwin, 2010) corpora comprising of 244 scientific papers classified into four of the ACM CCS 1998 categories was used to validate our tool. Hence, we applied a clustering algorithm to group semantically related papers and compared the generated clusters with the initial group assignments. The classification process makes use of CNA to create a discourse representation that facilitates the extraction of keywords and automated text categorization. The tool allows the development of categorization systems based on semantic similarity scores that are computed using semantic models such as LSA or LDA.

The dataset contained articles from different scientific fields categorized into four preselected 1998 CCS disjoint categories. A *k-means* clustering (Wu et al., 2008) was performed to classify the papers into four clusters, which were mapped to one of the initial four categories. Cohesion scores were used as indicators of the relevance of the papers within each cluster. The four resulting clusters were further mapped to the initial four categories covered by the scientific papers. Two experiments were performed: one using LSA and the other one relying on LDA, both models being pre-trained on collection of papers. A 4-means clustering was performed. Table 1 shows that for LSA the accuracy was 79%, while LDA correctly clustered 74% of the papers. Compared to the random chance of 25%, the tool provided valid results.

Table 1. Correctly-assigned papers per cluster and overall

Cluster	**Matching category**	**LSA**			**LDA**		
		Papers in cluster	**Correctly matched**		**Papers in cluster**	**Correctly matched**	
			#	**%**		**#**	**%**
1	*Distributed Systems*	40	38	95%	80	53	66.25%
2	*Information Search and Retrieval*	63	62	98.41%	68	64	94.18%
3	*Distributed Artificial Intelligence*	51	54	94.44%	61*	32*	52.46%
4	*Social and Behavioral Sciences*	90	38	42.22%	35	31	88.57%
	Total	244	192	**78.69%**	244	180	**73.77%**

4.3 Case Study 2: Text Categorization using Keywords

The previous case study showed that categorization systems based on textual contents of documents represented through discourse is of great help having an average accuracy of about 76-77%, depending on the implied semantic model. While this process is of great help when relying with a large collection of documents that should be classified, the computational power required increases exponentially with the number of documents. Consequently, the idea of transposing this classification system with one that relies only on keywords arose. An enhanced version of the *Keywords Extraction* tool was implied rather than relying on authors' keywords. Two experiments were performed: one using "unfiltered" keywords, either simple words or bigrams, and one using "filtered" keywords. *Filtered* keywords is based on an approach that "penalizes" words that are artificially emphasized by their number of occurrences but they do not really express so much value (common words). The same SemEval-2010 task 5 corpus (Popescu & Strapparava, 2015) consisting of 244 scientific papers belonging to four disjoint categories was used.

Initially, the keywords of the entire collection of papers were extracted using the initial version of the *Keywords Extraction* tool. The LSA semantic model was used for this experiment, pre-trained on the corpus containing the set of papers. New text files were generated with the extracted keywords, which were either simple words or bigrams. The bigrams were obtained using the Stanford CoreNLP (Manning et al., 2014) dependency parser. After that, a clustering based on the newly created text files was performed. The results can be seen in Table 2. It can be observed that for any case the percentage of correctly assigned papers was about 67-68%.

The following experiment considered only relevant words for the analysis. This approach requires the computation of the adjusted score by determining the positions of the words' occurrences. To obtain them, we built a custom text which consisted of the entire collection of papers and performed a keyword extraction on the aggregated text. The resulted keywords, which were either simple words or bigrams, were used to determine their "commonness" score. The list of keywords of the corpora was sorted by relevance. Of these we kept the most relevant keywords that made up to 10% of the maximum relevance, which lead us to 6,400 keywords, of which 805 were simple words, while 5,595 were bigrams.

Table 2. Correctly assigned papers without keyword filtering

Category	Papers in category	Matched cluster	Correctly classified	Percentage
Distributed Systems	59	2	50	84.75%
Information Search and Retrieval	64	3	63	98.44%
Distributed Artificial Intelligence	60	4	12	20%
Social and Behavioral Sciences	61	1	40	65.57%
Total	244		165	**67.62%**

Further, we applied the ALD formula to adjust the relevance so that words that have the occurrences very close one to another to be penalized, while words that are better spread across the corpora to be rewarded. Of the keywords that had the highest adjusted relevance we chose the first half as accepted keywords. A second clustering was performed using the keywords from the papers, but we kept only the keywords that appeared in this list. The percentages of correctly assigned papers per category and overall are presented in Table 3. It can be observed that the total percentage was about 78%, higher with more than 10% in contrast to the clustering performed without keyword filtering. This demonstrates the keyword extraction mechanism, enhanced with the algorithm of removal of common words as a valid tool to gather keywords of a text. Thus, by narrowing the set of words increase the accuracy of the tool increased by more than 10%.

Table 3. Correctly assigned papers with keyword filtering

Category	Papers in category	Matched cluster	Correctly classified	Percentage
Distributed Systems	59	3	53	89.83%
Information Search and Retrieval	64	1	63	98.44%
Distributed Artificial Intelligence	60	2	16	26.67%
Social and Behavioral Sciences	61	4	59	96.72%
Total	244		191	**78.28%**

4.4 Case Study 3: Quality Assessment for French CVs

This experiment was aimed at presenting a new tool built to support candidates in increasing the quality of their CV for a job. Both the visual quality and the textual content are considered while also providing an overview and corresponding feedback for the entire CV. The presented CV analysis tool uses advanced NLP techniques to interpret and understand the content from written texts, while also considering their visual traits. The study was performed on a collection of 52 CVs manually annotated as positive or negative in terms of their visual and content-oriented aspects. A statistical analysis was performed on more than 400 factors to extract the traits that define a good commercial CV. A custom tool was developed and integrated with the already available *ReaderBench* framework (Dascalu, 2014; Dascalu, Dessus, Trausan-Matu, Bianco, & Nardy, 2013; Dascalu et al., 2015a). The CVs were previously annotated as positive or negative in terms of their visual and textual content-oriented aspects.

While relating to the indices considered for subsequent analyses, we started by considering statistics regarding the structure of the text, i.e., the number of pages, paragraphs, sentences, words and content words. Visual aspects covered statistics like the number of images contained within the CV and the number of colors; both were normalized to the number of pages. Font statistics was another visual aspect considered; it included the number of font types, basic font types and font sizes used in the texts. Font sizes were relevant while relating to the number of different sizes, normalized by the number of pages. Statistics regarding the usage of Bold, Italic, and both Bold and Italic characters were performed, as well. Thus, the total number of corresponding characters was computed and normalized to the total number of characters. Words' valences were determined (such as positive, negative or neutral words) using the valence FAN scores. The number of words contained in categories of the LIWC list were calculated. Textual complexity indices available for French language were also computed and cover the following categories: surface, lexical, syntax, semantics, and discourse structure. The detailed presentation of the indices is available in previous work (Dascalu, 2014; Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014).

Two multivariate analyses of variance (MANOVA) (Garson, 2015) were conducted to examine the effect of each index in terms of the considered criterion. With regards to *visuals* aspects, the indices from Table 4 present (upper part), in descending order of effect size, the visual and surface indices that were significantly different.

Table 4 (lower part) presents, in descending order of effect size, the indices that were significantly different for content-centered characteristics. Two stepwise Discriminant Function Analyses (DFAs) were performed. The first DFA retained one variables as significant (*Simple font types*). The DFA had an accuracy of 63.5% (chance level = 50%). The second stepwise DFA, centered on predicting the quality of a CV's content, retained one variables as significant (*Adverbs LIWC*) and had an accuracy of 67.3% (chance level = 50%).

Table 4. Tests of between-subjects effects for significantly different indices predictive of the visual aspect

Index	***M (SD) positive***	***M (SD) negative***	***F***	***p***	**Partial η^2**
Visual aspects					
Simple font types	3.58 (1.34)	2.43 (1.72)	7.373	.009	.129
Minimum font size	8.51 (3.19)	6.38 (3.65)	4.711	.035	.086
Structural aspects					
Number of adverbs (LIWC)	11.22 (4.73)	6.92 (3.79)	12.966	.001	.206
Number of words labeled with positive emotions (LIWC)	9.11 (3.62)	6.68 (2.85)	7.155	.010	.125
Average words per sentence labeled with achievement (LIWC)	0.37 (0.17)	0.26 (0.13)	7.003	.011	.123
Sentece standard deviation in terms of unique words	5.00 (2.35)	3.52 (1.60)	6.909	.011	.121
Word entropy	4.74 (0.28)	4.57 (0.27)	5.447	.024	.098
Document flow average cohesion using path length similarity and maximum value criteria	0.44 (0.05)	0.40 (0.06)	5.014	.030	.091
Average number of syntactic dependencies per sentence (multi-word expression)	1.36 (0.57)	1.04 (0.45)	4.816	.033	.088
Average words per sentence labeled with leisure (LIWC)	0.23 (0.10)	0.17 (0.08)	4.741	.034	.087
Average number of syntactic dependencies per sentence (determiner)	1.80 (1.04)	1.28 (0.64)	4.688	.035	.086
Number of words labeled with inclusion (LIWC)	1.30 (0.87)	0.76 (0.93)	4.645	.036	.085
Average words per sentence labeled with friends (LIWC)	0.12 (0.07)	0.08 (0.06)	4.613	.037	.084

4.5 Case Study 4: Providing Support in MOOCs

This study was aimed at describing a prototype of a Massive Open Online Course (MOOC) platform that integrates various functionalities using automated NLP techniques to promote self-regulated learning. Discussions of the value of having widely available, accessible and flexible resources in higher education are also covered. Standard e-learning systems, such as Moodle (Dougiamas & Taylor, 2003) or BlackBoard, suffer from several problems that hamper the use of their entire functionalities. Although promoting themselves as providing "open" access, their content can be accessed through an enrollment. This provides a slower access to a course and makes the students' activity subject to analysis (by researchers, teachers and sometimes even by companies) without their agreement or sometimes without being aware of that. Moreover, each system has its own functioning and structuring, which forces teachers and students to adapt to them. The solution proposed in this research is twofold. On one hand, open access and flexible courses are promoted, in which learners are able to follow "personalized learning trajectories for themselves" (Brand-Gruwel, Kester, Kicken, & Kirschner, 2014, p. 363). On the other hand, using the recent advances in Natural Language Processing a layer leading to immediate interactions and questioning students is implemented. Thus, the proposed approach outlines the term of Massive Online Open Textbook (MOOT) promoted by Baker[7].

Self-regulated learning (SRL) occurs through four phases (Winne, 2011): defining the task, setting goals and plans, engagement and adaptation on a large scale. The student is accompanied by the teacher and by the computerized systems throughout these phases. By using the recent advances in NLP, various features that can help each one of these steps were implemented into a set of learning materials to demonstrate the functionality:

1. Defining the Task. The student may be allowed to perform more advanced searches than simple keywords by analyzing all the materials of a course with the help of information retrieval techniques like LSA and LDA.

2. Setting Goals and Plans. The student may be provided a concept map of the words appearing in the current learning material. The concept map may also display "inferred" words (i.e., words that are very close to the ones existent in the learning material).

[7] http://www.columbia.edu/~rsb2162/bigdataeducation.html

3. Engagement. By allowing students to rephrase a learning material with her own words can be of great help by allowing her to assess her current knowledge.

4. Adaptation on a Larger Scale. The MOOT platform could offer a forum space, where contributions could be automatically analyzed by comparing them to the whole discussion.

Figure 1 shows a course page within the learning material that integrates a concept map generated through a cached request (to avoid overloading the server) to the *ReaderBench* web server. The concept map shows the main keywords of the learning material and was generated with the use of LSA, LDA and WordNet trained on the *Le Monde* corpora. The nodes of the map can be rearranged with the help of the mouse to better visualize the words closest to a node. The size of a node is proportional to the keyword's importance in the learning material.

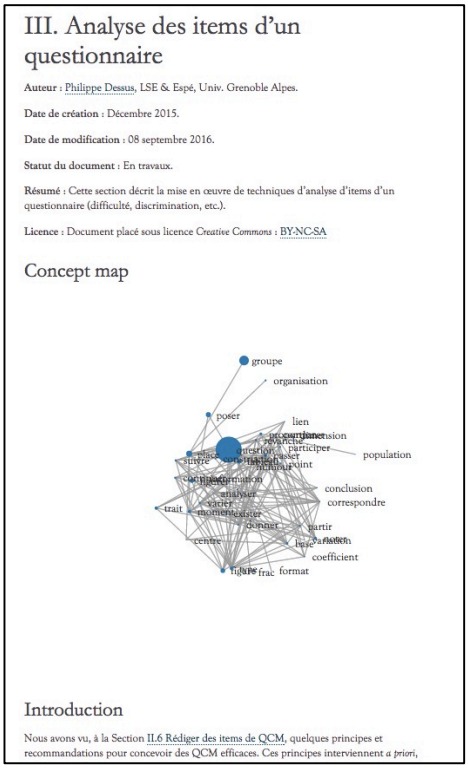


Figure 1. Screenshot of a MOOT incorporating a Concept Map generated through ReaderBench

Besides teacher assistance, the students may benefit from information retrieval techniques that consider the semantic characteristics of the keywords by using the implemented search functionality. The students may also gather a concept map of a learning material to help them to set more specific goals. Integration in the MOOT of a forum and the implementation of an automated analysis of the contributions would allow the students to have a better understanding of the topics by using a collaborative construction of knowledge (Stahl, 2006).

4.6 Case Study 5: Extraction of E-Learning Topics in a MOOC platform

This study was aimed at classifying a collection of learning materials for a MOOC platform created for the Early Nutrition eAcademy (ENeA) project. The envisioned optimal scenario considers the adaptation of the learning environment to national and regional needs of the participants, which can be very diverse. Mass customization is an economical concept aimed at enhancing flexibility and personalization of products, corroborated with lowering the costs of mass production. This concept can be also applied to education. This case study describes the technical solution for an automated analysis of learning needs based on the responses provided by a target group of users. The automated content analysis tool available within the *ReaderBench* framework was employed to extract and cluster key concepts extracted from the participant free text responses provided to an online questionnaire. Participants' learning interests were identified, which enabled the ENeA project to provide customized learning contents for a very high number of participants.

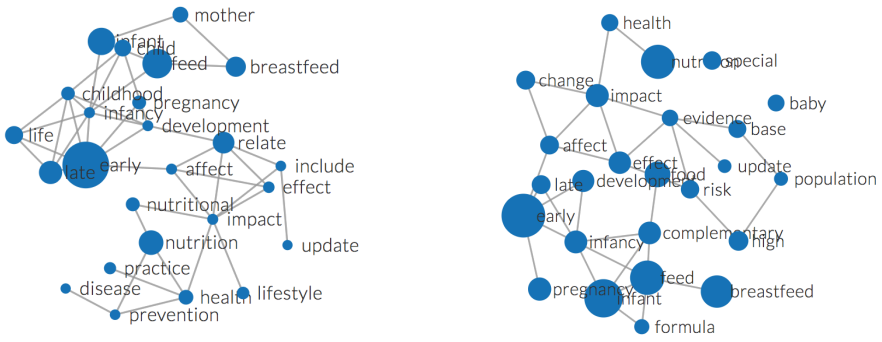
Continuing Medical Education (CME) in Early Nutrition is challenged by the diversity of participants (doctors in practice from different specialties and work settings, dieticians, nurses, medical students, etc.) with different knowledge levels and facing with regionally diverse problems (e.g., obesity in developed vs. malnutrition in developing regions). Such challenges call for a customized CME that, at the same time, can be largely available at affordable costs. To achieve this, the approach of mass customization (MC) (Pine, 1993) has been chosen for further ENeA development. Based on an in-depth customer (or learner) needs analysis, MC combines mass and customized production (or generation of individual online courses) so that it fulfills individual needs, while saving costs. MC applications in education requires clear delimitations from related concepts. Probably the most pervasive related concept is "adaptive and personalized learning environments" (Kinshuk, 2016). These environments centrally include a learner model that describes individual cognitive features such as the previous knowledge level, learning styles and preferences etc. In contrast, MC is mainly based on task models describing productive operations and their distribution across the organizational structure of the producer.

Prior to the experiment, *ReaderBench* was trained on a text corpus consisting of 1,700 specialized documents organized in 7 modules related to pregnancy, nutrition, epigenetics, and nutrients, which was combined with the TASA corpus (http://lsa.colorado.edu/spaces.html).

The network of ENeA project partners from a regional ENeA subproject located in Malaysia and Thailand were asked the question "What would you like to learn and be continually educated for in the field of Early Nutrition & Lifestyle?" in an online survey. For the free text responses provided by the entire participant sample, a total of 290 keywords were extracted, from which the most relevant five keywords extracted were: early (relevance score 3.97), feed (2.41), late (2.31), infant (2.30), and nutrition (1.96).

To find the minimal number of profession groups according to the importance of the extracted scores, the initially indicated groups were considered as cases described by keywords. A Principal Component Analysis with varimax rotation and Kaiser normalization was performed. Two factors resulted. For the profession group lecturer/nutritionist/pediatrician, a total of 237 keywords were extracted, from which the most relevant keywords were: early (3.69), feed (2.34), infant (2.16), nutrition (1.94), late (1.83). For the profession group dietician/doctor-in-practice, a total of 82 keywords were extracted, from which the most relevant keywords were: early (relevance score 1.71), infant (1.49), feed (1.34), nutrition (1.33), breastfeed (1.26).

Based on these distributions of keywords per profession group, the Early Nutrition expert, head of the ENeA project, indicated that: cluster 1 (left-hand side of Figure 2) focuses more on broad and general aspects of pre- and postnatal nutritional programming of long-term health, which may be interest for all health care professionals, students and lecturer. In contrast, cluster 2 (right-hand side of Figure 2) is centered on infant feeding, which may be of particular interest for pediatricians, but also for general practitioners (GPs) and other professional groups counseling families with infants.



a) b)

Figure 2. Early Nutrition Concept Map for the two identified interest clusters

5 Automated Detection of Implicit Links

5.1 Overview

This section covers the experiments conducted with regards to discourse analyses throughout chat conversations. The collection of chats consisted of conversations between students from our faculty. Different formulas were compared so that to determine whether the "raw" semantic scores require any adjustment. Experiments were also aimed at determining the best window to look for implicit links both in terms of number of utterances and in terms of time spent between posting times.

As Figure 3 shows, each conversation consists of utterances, while each utterance is comprised of words. Given this input data, the *Conversation Analysis* module analyses utterances to detect implicit links by using semantic models and WordNet ontologies. The utilized semantic models are LSA, LDA and Word2vec, while WordNet-related algorithms will be presented in the following section. The *Conversation Analysis* module detects implicit links by using different formulas and looking for references through various window sizes. The window sizes use two dimensions: the distance of utterances and the time passed between two postings.

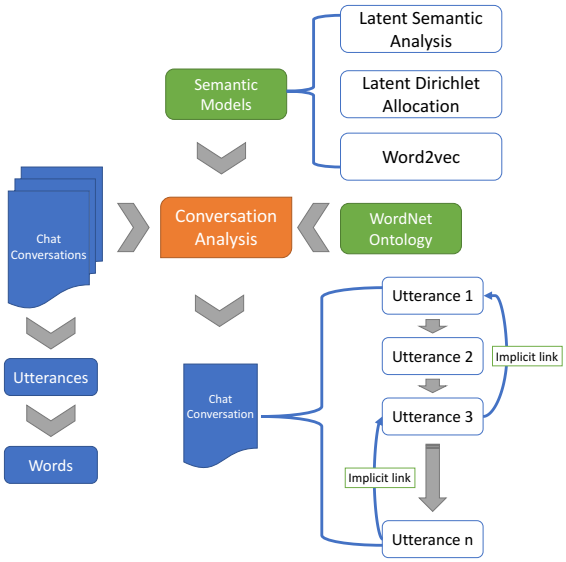


Figure 3. CSCL Conversation Analysis Using Semantic Models and WordNet for Detection of Implicit Links

5.2 The Training Corpora

The experiments involved semantic models that were trained on a custom corpus. It consisted of TASA (Touchstone Applied Science Associates, Inc., http://lsa.colorado.edu/spaces.html) – a large corpus that contains a variety of general texts, novels and newspaper articles – for general words. For the statistical methods, we had to choose one corpus that covered most words expressed by chat participants. Of these words, some of them, like "forum", "chat" or "blog" have been introduced very recently into the English language. While some of the corpora that demonstrated maturity in that they have been widely studied within the latest years were published around the beginning of 1990s (TASA is 28 years old at the time of writing this thesis), such words did not occur as much as expected or they expressed a different meaning. For example, the word "forum" nowadays mostly refers to an internet forum, a collaborative online discussion platform were people hold conversations with regards to specific topics. Prior to the modern era, "forums" referred to public squares in the Roman empire, which were used as marketplaces, for social discussions or activities, meetings and others. Our LSA, LDA and Word2vec models were trained on a pre-processed version of a custom corpus obtained by concatenating the TASA corpus and a corpus of more than 500 CSCL-related scientific papers.

5.3 The Corpus of Conversations

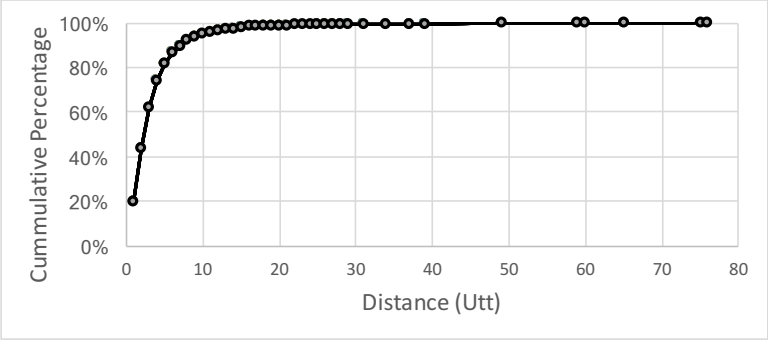
The collection of chats involved in our CSCL studies consisted of conversations performed by computer science students from our faculty using the *ConcertChat* environment (Holmer, Kienle, & Wessner, 2006). This application allows users to explicitly reference previous utterances – we call these *explicit links* or *explicit references*. The main topic of the discussions was the emphasizing of the benefits and disadvantages of each of the several web collaboration technologies (i.e., wiki, blog, forum, chat) and to identify the most suitable tool to be used by an enterprise (Trausan-Matu & Rebedea, 2010). While showing contrasting opinions, the students were asked to reach an agreement at the end of the deliberation.

A collection of 55 chat conversations was used. The collection totalized 17,612 utterances with an average number of 4.35 participants. The total number of explicit references was 4,463, while the average coverage (i.e., the percentage of referred utterances by the total number of utterances) was 28.62%. The average time duration of a conversation was about 2 hours.

Figure 4 presents the graphical evolution of the coverage of explicit links as a function of distance and time. We can observe in a visual manner that a window size of 20 utterances ensures the coverage of most (99%) explicit links, while a distance of 10 utterances enables a sufficient level of certainty (covering more than 95% of the total links). Given these distributions, we decided to compute the semantic similarity between each utterance and the previous ones considering window sizes of 20, 10 and 5 utterances. Regarding the time difference between utterances, several time frames were set and explicit references' coverage within these time frames were computed. Significant changes in terms of cumulative percentage were desired, which made us to select 5 time frames for the study: 30 seconds, 1, 2, 3 and 5 minutes. As it was observed, within 5 minutes 97% of the explicit references are covered, while the time frame of 1 minute covers 61% of them.

A sample conversation file is presented in **Appendix I**.

(a)



(b)

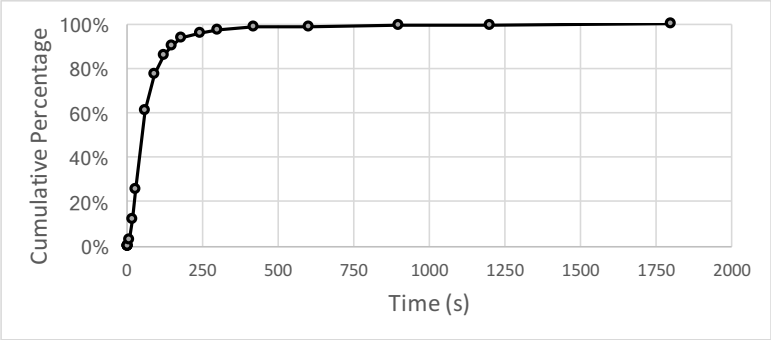


Figure 4. Cumulative coverage of explicit links (a) per distance and (b) per time

5.4 Case Study

After analyzing various semantic methods for assessing of semantic similarity between words, two experiments aimed at automated detection of implicit links in chat conversations were performed (Gutu, Dascalu, Rebedea, & Trausan-Matu, 2017; Gutu, Dascalu, Ruseti, Rebedea, & Trausan-Matu, 2017). Two accuracy measures were considered within these studies: a) *exact-match* implicit links detection, when the computed reference is the same as the explicit reference attribute previously set by the user, and b) *in-turn match* implicit links detection, when the computed reference belongs to the same turn, a collection of adjacent utterances belonging to the same participant, including the utterance mentioned within the explicit link. Our corpus of 55 chat conversations was initially cleaned using several NLP refinements (Manning & Schütze, 1999): stop-words (words with no semantic relevance and no contextual information) were eliminated, duplicate words frequently encountered in chat conversations were removed and the remaining words were lemmatized using the Stanford Core NLP library (Manning et al., 2014). This experiment had to purposes: 1) to determine which of the state of the art methods for computing semantic similarity performs best for the detection of implicit links in multi-party chat conversations and 2) to discover the optimum distance and time frame to look for implicit links. The corpus for this comparative analysis consisted of the cleaned collection of 55 chat conversations that lasted up to two hours.

Table 5 shows two examples of identified implicit links extracted from the same conversation. The examples show the differences between in-turn matching, when implicit links must belong to the same participant in a continuous block of utterances, and exact matching when implicit links must overlap perfectly with the explicit links defined by the user. The first example shows the identified implicit link for utterance 140 was turn 138 using the same parameters for time frame, distance and semantic similarity. Turn 138 was also the explicit reference for utterance 140, as it can be easily observed from the Ref. ID column, and this was a correct *exact matching*. In the second excerpt, the utterance with id 74 having an explicit reference to utterance 65 which was manually added by a participant (explicit links are in the second column – Ref. ID). However, when imposing a window of maximum 5 utterances and 1 minute time frame, the detected implicit link was utterance 72 (emphasized) using the Path Length similarity measure. As turn 72 is enclosed in a continuous series of utterances belonging to the same user (i.e., Monica), we considered this to be a correct *in-turn matching*, but an incorrect *exact matching*.

Table 5: Excerpt from a chat conversation showing an in-turn matching and an exact matching

Utt. ID	Ref. ID	Implicit link	Speaker	Time	Content
			Exact matching		
137			Monica	09:21:41	blogs are a good solution
138			**Stefan**	**09:22:01**	**you know the biggest disadvantage of wikis? that anybody can input and that makes wikis a not-so-reliable source of info**
139	138		Alex	09:22:20	that depends on the configuration
140	**138**	**138**	Razvan	09:22:24	you could have admins that check the information
			In-turn matching		
65			**Monica**	**09:08:27**	**features to add RSS feeds, file sharing and so on**
			... (several utterances of the same participant, Monica) ...		
72			**Monica**	**09:09:57**	**and they embed only what you need**
73			Monica	09:10:16	users tend to be scared away by a multitude of features that they need to figure out
74	**65**	**72**	Razvan	09:10:22	The thing that I think would be a problem with wikis is that they will not allow a person to keep track of the latest information added. Ok RSS are good but not everybody wants to use an RSS feed reader.

Three formulas were used: semantic similarity (SIM), normalized similarity by inverse distance between current utterance and referred utterance (NSIM), and semantic similarity computed using Mihalcea's formula (MSIM) (Mihalcea, Corley, & Strapparava, 2006). SIM represents the baseline similarity metric for each semantic model; for example, for LSA we use SIM to refer to the standard formula for cosine similarity. NSIM is used to refer to a normalized value of the previously introduced similarity. The formula developed by Mihalcea increases the similarity score of a pair of utterances by the highest similarity score between one word belonging to an utterance, and another belonging to the other utterance. Table 6 presents the percentage of detected explicit links using both exact and in-turn accuracy measures. Bolded values represent the formula that provided the best accuracy for each technique and for each windows size. Normalized semantic similarity provided the best accuracy for most of the used techniques and for all the three selected window sizes: about 30% for perfect match and about 40% for in-turn match.

Table 6. Implicit links detection rate per (window size, time frame) pair (exact matching / in-turn matching)

Window size, time frame	Measure	Leacock	Wu-Palmer	Path length
5 utterances 1 minute	SIM	**28.55%** / **8.18%**	**30.95%** / **40.44%**	**32.44%** / **41.49%**
	NSIM	26.78% / 36.19%	27.73% / 36.81%	31.35% / 40.51%
	MSIM	27.06% / 36.75%	28.89% / 38.45%	31.04% / 39.97%
10 utterances 1 minute	SIM	**27.82%** / **37.30%**	**30.27%** / **39.62%**	**31.88%** / **40.78%**
	NSIM	26.91% / 36.29%	27.88% / 36.94%	31.57% / 40.65%
	MSIM	26.16% / 35.59%	27.99% / 36.95%	30.60% / 39.16%
5 utterances 2 minutes	SIM	**28.55%** / **38.18%**	**30.95%** / **40.44%**	**32.44%** / **41.49%**
	NSIM	26.78% / 36.19%	27.73% / 36.81%	31.35% / 40.51%
	MSIM	27.06% / 36.75%	28.89% / 38.45%	31.04% / 39.97%
10 utterances 2 minutes	SIM	**27.82%** / **37.30%**	**30.27%** / **39.62%**	**31.88%** / **40.78%**
	NSIM	26.91% / 36.29%	27.88% / 36.94%	31.57% / 40.65%
	MSIM	26.16% / 35.59%	27.99% / 36.95%	30.60% / 39.16%
Window size, time frame	**Measure**	**LSA**	**LDA**	**Word2vec**
5 utterances 1 minute	SIM	30.68% / 39.72%	28.77% / 38.61%	**30.56%** / **39.36%**
	NSIM	30.01% / 38.74%	25.43% / 34.73%	25.32% / 34.64%
	MSIM	**31.45%** / **40.41%**	**30.98%** / **40.09%**	28.16% / 38.18%
10 utterances 1 minute	SIM	29.25% / 37.80%	27.12% / 36.27%	**29.46%** / **37.64%**
	NSIM	**30.28%** / **38.94%**	25.55% / 34.83%	25.37% / 34.71%
	MSIM	30.28% / 38.48%	**30.08%** / **38.67%**	26.81% / 36.16%
5 utterances 2 minutes	SIM	30.68% / 39.72%	28.77% / 38.61%	**30.56%** / **39.36%**
	NSIM	30.01% / 38.74%	25.43% / 34.73%	25.32% / 34.64%
	MSIM	**31.45%** / **40.41%**	**30.98%** / **40.09%**	28.16% / 38.18%
10 utterances 2 minutes	SIM	29.25% / 37.80%	27.12% / 36.27%	**29.46%** / **37.64%**
	NSIM	30.28% / **38.94%**	25.55% / 34.83%	25.37% / 34.71%
	MSIM	**30.28%** / 38.48%	**30.08%** / **38.67%**	26.81% / 36.16%

6 Extending the *ReaderBench* Services

6.1 Overview

Most of the experiments presented in this thesis have applicability for regular users and thus they were exposed through publicly available demos on the *ReaderBench* website (http://readerbench.com). Up until recently, the desktop version of the *ReaderBench* framework was hardly usable in hands-on educational contexts due to the requirements of extensive processing power and high amounts of memory usage. Due to these limitations, it was mostly used in follow-up offline analyses. The online version opens new usages of *ReaderBench* in education, as our framework can now be effectively used in a wide range of educational situations and needs. The desired outcome of exposing the services is to: 1) provide access to people in order to open borders in open learning and 2) allow researchers and developers create their own modules based on *ReaderBench*'s services through the free access Application Programming Interface (API).

6.2 The Semantic Annotation Tool

The Semantic Annotation tool automatically classifies documents in accordance to a pre-imposed list of categories extracted from the level 1 categories of the 2012 ACM CCS. The tool also extracts relevant concepts as potential keywords and analyses the semantic relatedness between the keywords of the paper, the abstract and the document (Gutu et al., 2017). The online tool is available on the *ReaderBench* website (http://readerbench.com/demo/semantic-annotation). One randomly-selected paper of the *SemEval* collection (Kim, Medelyan, Kan, & Baldwin, 2010) was selected. The article (Laskowski & Chuang, 2006) focuses on presenting a new network monitoring capability that should reflect a new economic model for Internet Service Providers (ISP). A preview of the paper and the corresponding outputs of the analysis performed using the online tool are presented in Figure 5. The *Keywords Map* generates a graph depicting the most relevant keywords related to the conceptualization of the paper as nodes, while edges reflect semantic links between the nodes above an imposed threshold. The size of a node is proportional to its relevance score and a threshold of 0.4 was used for selecting links above the imposed semantic similarity.

The paper was annotated with the *C.2.4 Distributed Systems* category of the 1998 CCS taxonomy. Our annotation tool, which relies on the 2012 version, provided the highest score for *Theory of computation*, which we consider adequate as the paper describes a new computational model. *Software and its engineering* is, also, a category with a high similarity score – adequate in our opinion as the paper describes the proposed system. *Computing methodologies* is the third category by relevance score, also meaningful in since the paper presents a new method to compute how much one individual should pay for their Internet connection based on their activity. The top 10 keywords of the paper include "*model*", "*order*" and "*design*" which outline the general idea of the paper. The *Keywords Overlap* section shows the number of occurrences and relevance scores for each of the authors' keywords – two out of the five keywords were found with similarity scores around 0.3. The *Relevance Scores* show how related are the abstract, the authors' keywords, and the whole document – this paper exhibits a high semantic similarity between the abstract and its content, but a lower one between the keywords and the paper motivated by the fact that authors specified only five keywords.

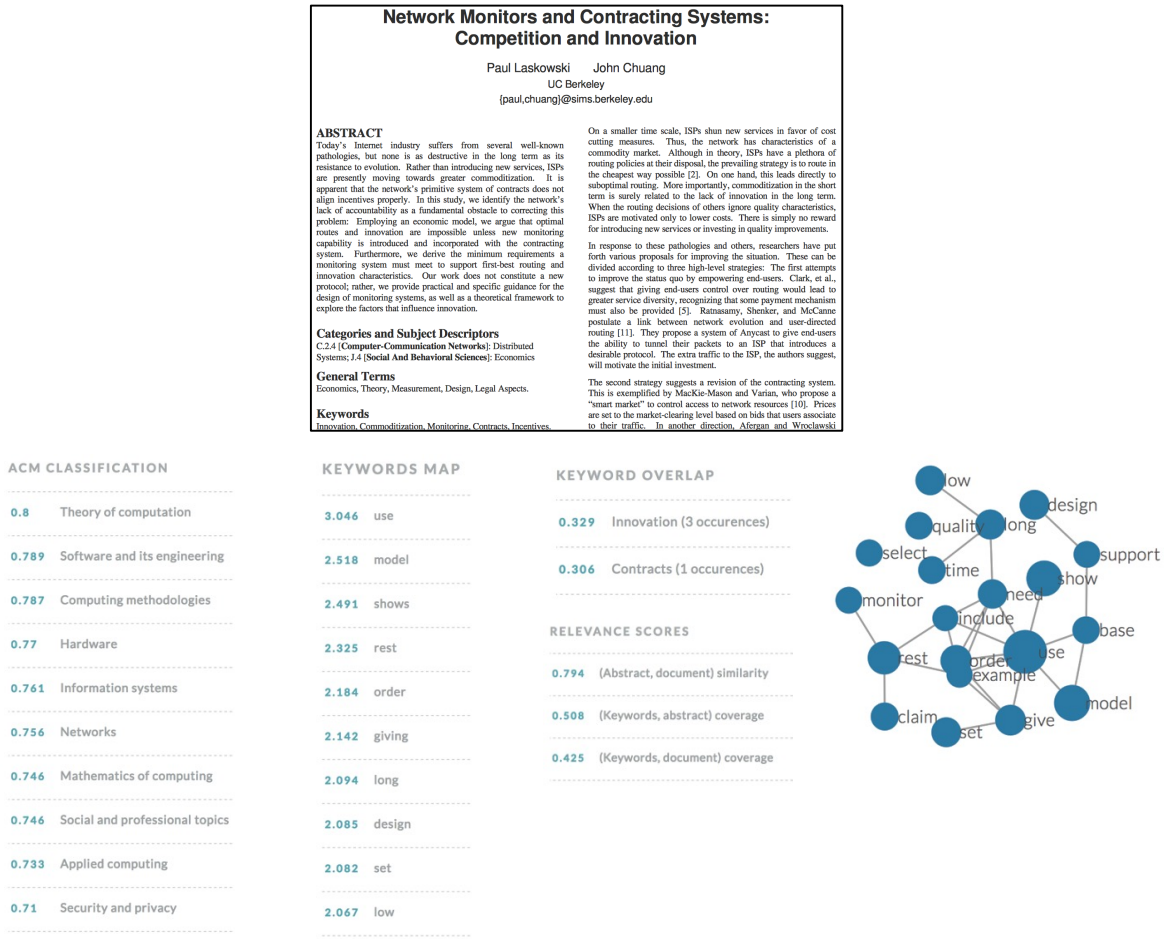


Figure 5. The Semantic Annotation tool's results for a sample paper (Laskowski & Chuang, 2006)

6.3 The Enhanced Keywords Extraction Tool

The enhanced version of the *Keywords Extraction* tool was validated through the experiment relying on a method of scoring "commonness" of words (Gutu, Ruseti, Dascalu, & Trausan-Matu, 2017). It integrates bigrams and "penalizes" words artificially scored as relevant because of a high number of occurrences through a mechanism of computing a "commonness" scores (http://readerbench.com/demo/keywords). Outputs for the sample paper are displayed in Figure 6. The most relevant keywords of the paper, either simple words or bigrams, are presented.

Word(s)	Lemma(s)	POS	Occ	Links	Degree	Relevance	TF	IDF	Similarity			Avg. Dist. to Root	Max. Dist. to Root	Polysemy Count	Related words
									LSA	LDA	Word2Vec				
new	new	JJ	3	8	4.433	1.261	2.099	2.454	0.473	0.728		0	0	9	▲ Hide related words • protocol new (NN JJ) - 1.000 • capability new (NN JJ) - 0.997 • service new (NN JJ) - 0.823 • industry today (NN NN) - 0.349 • today (NN) - 0.331 • impossible introduce (JJ VB) - 0.317 • great (JJ) - 0.309 • term long (NN JJ) - 0.307
system	system	NN	3	14	9.640	1.078	2.099	5.819	0.518	0.509		8	8	9	▲ Hide related words • align system (VB NN) - 0.997 • monitor system (VB NN) - 0.996 • system monitor (NN NN) - 0.996 • system primitive (NN JJ) - 0.992 • system network (NN NN) - 0.989 • system contract (NN NN) - 0.919 • support system (VB NN) - 0.905 • meet system (VB NN) - 0.885 • incorporate capability (VB NN) - 0.357 • network (NN) - 0.350 • model economic (NN JJ) - 0.327 • economic (JJ) - 0.312 • argue provide (VB VB) - 0.310 • capability monitor (NN NN) - 0.305
service_new	service new	NN JJ	1	15	7.091	0.647	1.000	2.454	0.528	0.765		6	6	11	▼ Show related words
meet_system	meet system	VB NN	1	18	10.579	0.633	1.000	5.819	0.563	0.703		4	4	11	▼ Show related words
route_innovation	route innovation	VB NN	1	6	3.776	0.617	2.000	458.619	0.263	0.355		7	7	3	▼ Show related words
capability_new	capability new	NN JJ	1	8	4.445	0.613	1.000	2.454	0.495	0.732		3.5	3.5	6	▼ Show related words
introducing	introduce	VB	2	8	5.078	0.605	1.693	51.488	0.283	0.432		0	0	10	▼ Show related words
protocol_new	protocol new	NN JJ	1	8	4.429	0.601	1.000	2.454	0.474	0.728		4	4	6	▼ Show related words
support_system	support system	VB NN	1	28	14.291	0.572	1.000	5.819	0.553	0.592		5.5	5.5	10	▼ Show related words
term_long	term long	NN JJ	1	19	7.553	0.557	1.000	3.374	0.412	0.703		2.5	2.5	8	▼ Show related words

Figure 6. The Keywords Extraction tool's results for the sample paper (Laskowski & Chuang, 2006).

6.4 The CV Analysis Tool

The CV Analysis tool allows users to gather recommendations regarding their CV written in French language and is available on the *ReaderBench*'s website through the demo section at http://readerbench.com/demo/cv. Visual displays for a sample CV are shown in Figure 7. The tool is meant for both regular users who wish to improve the overall quality of their CV, as well as for employers who can set their own list of keywords for the CVs to be sought for.

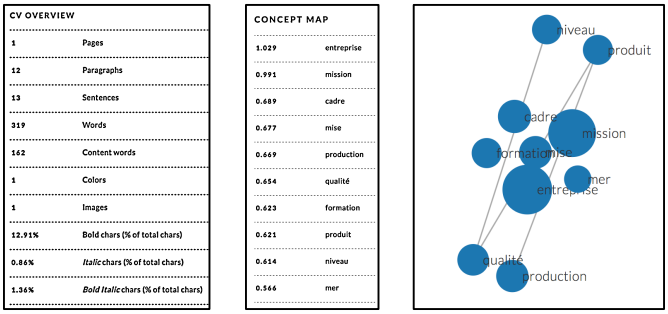


Figure 7. An example of output results for a sample CV

7 Discussions

7.1 Advantages of our Approach

This thesis describes the results obtained for two main tracks: detection of implicit links in chat conversations on one hand and discourse analysis for particular scenarios on the other hand. With regards to chat conversations, the performed comparative analysis provided evidence that the un-normalized semantic similarity measures and a window size of 5 utterances are the best trade-off in terms of exact and in-turn detection of implicit links. Regarding analysis of documents gathered from specific scenarios provided results with immediate applicability into many fields. Automated categorization can be performed either by relying on the full textual content of documents or by relying on its keywords, the second approach having a little lower accuracy, but saving a lot of required computational power. The analysis of CV documents has immediate applicability for both companies and for individuals; The presented experiments have applicability into education, too, by allowing teachers to integrate tools that help students in the learning process so that to set their goals and to follow them on one hand, but also to structure their courses on the other hand.

7.2 Faced Problems and Provided Solutions

The experiments presented in this thesis lead to few problems that were resolved as follows. With regards to the process of detection of implicit links in chat conversations, when relying on semantic models we had to choose the training corpus. While TASA is a collection of documents consisting of both scientific and regular articles, thus containing words from various domains, it is rather old, thus it doesn't contain terms covered throughout our collection of conversations with their modern meaning. Examples include "forum" or "chat", which in the modern era express a different meaning than in the past. Also, the old meaning is very seldom used nowadays, thus the terms couldn't be found that often. The implemented solution was to create a custom corpus consisting of both TASA and a collection of scientific papers related to CSCL, which contain many occurrences of words of our collection of conversations.

The experiment performed on the collection of CVs showed some challenges as the CVs did not follow a specific pattern, but were structured in many formats in terms of sections, number of columns, spaces between paragraphs and many others. The Apache PDFBox library was not able to correctly extract texts for any file, thus a minor adjustment to fix paragraphs was performed. Even so, the multiple types of CVs bring the necessity to perform more advanced "fixes" for issues with paragraph extraction.

The exposure of services through the API leads to an overwhelmed server, thus leading to much longer response times for both the API and the clients available through the website. Several optimizations in terms of memory were implemented through the code and physical adjustments like system upgrades were performed. Even though, the multiple services available makes *ReaderBench* a desirable solution, thus a long-term approach should be thought.

7.3 Educational Implications

The experiments presented in this thesis were focused on discourse analysis based on semantic models and textual complexity. The implication of modern semantic models and techniques together with classical textual complexity factors allowed us to gather valuable information from different types of documents. The integration of automated NLP techniques into the education, business and collaborative learning field could help into developing a more automated society that would not discriminate among individuals, but would give them equal opportunities to evolve. It could also allow the ones "left behind" to reach faster the "straight line" by providing them materials adapted to their current knowledge and disposition.

The obtained results for CSCL chat conversations open access for integration of a wide variety of facilities. While chat popularity increases[8] nowadays, efforts to adapt the supporting software applications in that to not only facilitate collaboration, but also to extract the most relevant information of them, should be the focus. The scientific community does not rely on a "standardized" application for chat conversations aimed at providing CSCL processes. With the popularity of modern applications increasing, a transition of the researchers' discussions to applications that provide the ability to create groups was observed (Soller, 2001). Efforts to

[8] http://www.icmi.com/Resources/Chat/2015/11/The-Stats-Behind-Chat-Its-Popular-and-Growing

create modules for such messaging apps or to create additional apps that use such protocols, like the Signal protocol (Ermoshina, Musiani, & Halpin, 2016), should be considered.

With regards to discourse analysis of documents, the models used throughout the experiments could be applied on a wide variety of texts with minimum adjustment efforts. As showed that classification of collection of documents could be completely automated by providing a set of categories and relying on text, this may have applicability in many domains. It would be interesting to take these models and use them for experiments in many research areas to discover whether there are any particularities involved that are necessary for maintaining an accurate model. A more advanced idea could be the development of an application that, based on some inputs of the person's education and experience, could lead to an automated generation of the person's CV. By adding an additional parameter consisting of the description of the job, the CV may be even adapted to match that specific job aimed at covering the company's requirements. With regards to the studies performed on MOOC platforms, the teachers could focus on creating the learning materials without the effort of thinking where would it fit, but simply provide it "as it is". The integration of more advanced service like automated assessment of students' deliverables (like their homework) would again ease the teachers' efforts by allowing them to automatically score students. Although, the teachers' effort would focus on creating materials that fit different levels of education by such an approach that allows every student to learn even if they lack some previous required education skills. These automated processes would provide a benefit for all the students by providing them equal chances by adapted learning materials and uniform scoring.

Regarding the learning process, our experiments have applicability also in education scenarios that involve children. Thus, by extracting similar words starting from a seed in a recursive manner, children can figure out how can they use words that they didn't know before. The linkage between the two words would allow them to see the connection, but also to see the slight difference that particularize the new terms. The accessibility of teaching resources and their flexibility is an increasingly important criterion for use by teachers and students. Their accessibility allows their use on many different media (tablets, computers), and their flexibility makes them usable in many scenarios: class feedback, inverted class, hybrid courses, continuous evaluation, self-learning, etc.). Moreover, it allows the exercise of a formative evaluation of the type "shoring", supporting the processes of self-regulation of the learner via these different tools (Allal & Mottier-Lopez, 2005).

8 Conclusions

8.1 Personal Contributions

The experiments presented in this thesis brought contributions into the NLP field, more specifically mainly with regards to discourse analysis based on semantic modelling and textual complexity. These covered processes of detection of implicit links in chat conversations and mechanisms of discourse analysis on different types of documents. Open access to the framework was also provided through an API with free access. The contributions are detailed below, structured on their main topic of research.

Detection of Implicit Links in CSCL chat conversations:

- Compared semantic methods (ontology-based and semantic models) and showed that path length relying on WordNet provided the highest accuracy for both exact match (32.44%) and in-turn match (41.49%).
- Discovered the optimal window to look for implicit links: 5 utterances and a time frame of 1 minute.
- The most accurate formula for the detection of implicit links was the un-normalized similarity score provided by each semantic method.

Discourse Analysis on different types of documents:

Categorization of scientific papers:

- Automated categorization performed on their content obtained an accuracy of 78.69% for LSA and 73.77% for LDA.
- Automated categorization performed on specific keywords (keywords that were too common were removed) obtained an accuracy of 78.28% for LSA.
- Showed that LSA method provided the best accuracy for categorization performed on content.
- Showed that, by limiting from full texts gathered from a paper to the list of specific keywords the accuracy losses only 1% for LSA, but the required computational power is much smaller.

Analysis of French CV:

- Showed that the judgement of the visual aspects of a CV is given by the number of font types that are used and the minimum size of the font.
- Showed that the judgement of the textual contents of a CV is given mostly by the number of adverbs, positive emotions, achievement words, standard deviation of sentences in terms of unique words and word entropy.
- Demonstrated that the judgement of the quality of a CV in terms of visual aspects and textual contents could be automated to ease the recruiters' job.

Analysis of MOOC platforms:

- Showed that the integration of discourse analysis facilities could help students of a courses platform to set their goals and engage to the learning process through keywords extraction, generation of concept map or automated assessment of their understandings.
- Showed that classification of learning materials into most suitable categories could be used for automated creation of learning modules through a case scenario related to the early nutrition domain in medicine.

Supported the development of additional ReaderBench features and functionalities:

- Created the *ReaderBench* website together with documentations to support installation and usage of available services.
- Integrated the tools used in the experiments into the *ReaderBench* website through demo clients to allow usage for regular users: semantic annotation, keywords extraction, CV analysis.
- Supported the development of *ReaderBench* with help in integrating of additional languages: setting up the dictionaries, integration of stop words, construction of (lemma, word) pairs, integration of pronouns and connectives.
- Supported the development of *ReaderBench* with help in training more corpora: TASA, TASA & CSCL, Le Monde (see dedicated experiment for details about these corpora).
- Transposed other *ReaderBench* facilities into website demo clients: sentiment analysis, reading strategies extraction, textual complexity assessment, CSCL analysis.
- Integrated all the services into the open-source *ReaderBench* framework.
- Provided open access for other researches without the need to install the *ReaderBench* framework through the integrated Application Programming Interface.

8.2 Directions for Future Research

With regards to chat conversations, although the thesis showed significant results, the detection of implicit links requires more investigations. Adjustments could be implemented to increase the accuracy of our identification process. First, machine learning techniques could be used to create an aggregated similarity score relying on multiple semantic measures. Second, dynamic sliding windows could be enforced by considering cut-offs induced by topic changes or long pauses within the discourse. Third, certain patterns extracted using speech acts (e.g., continuations, question answering) (Searle, 1969) and discourse connectors may be indicative of implicit links within the discourse.

For discourse analysis performed on different types of documents, a usage scenario for the Semantic Annotation tool consists of its integration within the RAGE[9] (Realising an Applied Gaming Eco-system) project to facilitate the automated classification of the publications contained in the internal DL. The service could be enhanced by a machine learning algorithm that would build and automatically annotate documents of a specific corpora in classes that will be fine-tuned with every step to define a specific classification for the corpora. Moreover, a comparison between the list of extracted keywords and the authors' keywords could be performed. Envisioned enhancements for CV analysis cover the usage of a larger dataset and the consideration of characteristics like age, location or gender to determine particularities of demographic groups. Another potential research targets the creation of a collection of representative English CVs. For the integration of facilities relying on NLP into course platforms analyses of the impact for either student' engagement or for the improvement of their results could be performed. For automated classification of learning materials into modules, different levels of granularity and different classification schemas could be tested to determine which one provides more engagement or is easier to be used.

By using the exposed services through the API, researchers and developers could perform their experiments while relying to data interpreted by *ReaderBench*. As the experiments were implemented into the afore mentioned open-source framework, development of additional studies could be performed.

[9] http://rageproject.eu

List of Publications

1. Rebedea, T., & Gutu, G. M. (2013). Detecting Implicit References in Chats Using Semantics. In D. Hernández-Leo, T. Ley, R. Klamma & A. Harrer (Eds.), *Scaling up Learning for Sustained Impact: 8th European Conference, on Technology Enhanced Learning, EC-TEL 2013, Paphos, Cyprus, September 17-21, 2013. Proceedings* (pp. 627-628). Berlin, Heidelberg: Springer Berlin Heidelberg. (ISI Web of Knowledge)
2. Rebedea, T., Chiru, C. G., & Gutu, G. M. (2014). How useful are semantic links for the detection of implicit references in CSCL chats?. In *RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, 2014* (pp. 1-6): IEEE. (ISI Web of Knowledge)
3. Gutu, M. G., Rebedea, T., & Trausan-Matu, S. (2015). A comparison of semantic similarity techniques for a corpus of CSCL chats. In *RoEduNet International Conference-Networking in Education and Research (RoEduNet NER), 2015 14th* (pp. 178-183): IEEE. (ISI Web of Knowledge)
4. Gutu, G., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2016). ReaderBench goes Online: A Comprehension-Centered Framework for Educational Purposes. In A. Iftene & J. Vanderdonckt (Eds.), *Romanian Conference on Human-Computer Interaction (RoCHI 2016)* (pp. 95–102). Iasi, Romania: MATRIX ROM. (BDI)
5. Gutu, G., Dascalu, M., Rebedea, T., & Trausan-Matu, S. (2017). Time and Semantic Similarity – What is the Best Alternative to Capture Implicit Links in CSCL Conversations? In *12th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2017)* (pp. 223–230). Philadelphia, PA: ISLS. (Category A in CORE 2017 Conference Ranking)
6. Gutu, G., Dascalu, M., Ruseti, S., Rebedea, T., & Trausan-Matu, S. (2017). Unlocking the Power of Word2Vec for Identifying Implicit Links. In *17th IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2017)* (pp. 199–200). Timisoara, Romania: IEEE. (BDI, Category B in CORE 2017 Conference Ranking)
7. Gutu, G., Dascalu, M., Heutelbeck, D., Hemmje, M., Westera, W., & Trausan-Matu, S. (2017). Semantic Annotation and Automated Text Categorization using Cohesion Network Analysis. In *5th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 13th Int. Conf. on eLearning and Software for*

Education (eLSE 2017) (pp. 25–32). Bucharest, Romania: Advanced Distributed Learning Association.

8. Gutu, G., Dascalu, M., Trausan-Matu, S., & Lepoivre, O. (2017). How Adequate is your CV? Analyzing French CVs with ReaderBench. In *3rd Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2015), in conjunction with the 21th Int. Conf. on Control Systems and Computer Science (CSCS21)* (pp. 559-565). Bucharest, Romania: IEEE. (BDI)
9. Gutu, G., Ruseti, S., Dascalu, M., & Trausan-Matu, S. (2017). Keyword Mining and Clustering based on Cohesion Network Analysis. In D. Trandabat & D. Gifu (Eds.), *3rd Workshop on Social Media and the Web of Linked Data (RUMOUR 2017), in conjunction with the Joint Conference on Digital Libraries (JCLD 2017)* (pp. 23–30). Toronto, Canada: "Alexandru Ioan Cuza" University Publishing House.
10. Dessus, P., Gutu, G., Dascalu, M., Diouf, J. B., & Trausan-Matu, S. (2017). Vers des manuels de cours universitaires ouverts et interactifs promouvant l'apprentissage auto-régulé. In *Atelier "Evaluation formative pratiquée en classe ou en amphithéâtre" joint à l'ORPHEE*. Font-Romeu, France.
11. Dascalu, M., Gutu, G., Paraschiv, I. C., Ruseti, S., Dessus, P., McNamara, D .S., Crossley, S., & Trausan-Matu, S. (2017). *Cohesion-Centered Analysis of CSCL Environments using ReaderBench*. Paper presented at the 18th Int. Conf. on Artificial Intelligence in Education (AIED 2017) – Interactive Event, Wuhan, China. (Category A in CORE 2017 Conference Ranking)
12. Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D .S., Crossley, S., & Trausan-Matu, S. (2017). ReaderBench: A Multi-Lingual Framework for Analyzing Text Complexity. In *12th European Conference on Technology Enhanced Learning (EC-TEL 2017)* (pp. 495-499). Tallinn, Estonia: Springer, Cham. (BDI)
13. Nistor, N., Dascalu, M., Gutu, G., Trausan-Matu, S., Choi, S., Haberman-Lawson, A., Brands, B., Korner, C., & Koletzko, B. (2017). Mass Customization in Continuing Medical Education: Automated Extraction of E-Learning Topics. In *12th European Conference on Technology Enhanced Learning (EC-TEL 2017)* (pp. 576-579). Tallinn, Estonia: Springer, Cham. (BDI)

14. Gutu, G., Paraschiv, I. C., Dascalu, M., Cristian, G., & Trausan-Matu, S. (in press). Analyzing and Providing Comprehensive Feedback for French CVs with ReaderBench. *Scientific Bulletin, University Politehnica of Bucharest, Series C*.
15. Radoi, I., Gutu, G., Rebedea, T., Neagu, C., & Popa, M. (2017). Indoor Positioning inside an Office Building Using BLE. In *Control Systems and Computer Science (CSCS), 2017 21st International Conference on Control Systems and Computer Science (CSCS21)* (pp. 159-164). IEEE. (BDI)

References

Allal, L., & Mottier-Lopez, L. (2005). L'évaluation formative de l'apprentissage : revue de publications en langue française *L'évaluation formative. Pour un meilleur apprentissage dans les classes secondaires* (pp. 265–299). Paris: OCDE.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin and London: The University of Texas Press.

Bestgen, Y. (2012). Évaluation automatique de textes et cohésion lexicale. *Discours, 11*. doi: 10.4000/discours.8724

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Bond, Francis, & Foster, Ryan. (2013). Linking and Extending an Open Multilingual Wordnet. In *ACL* (pp. 1352-1362).

Brand-Gruwel, S., Kester, L., Kicken, W., & Kirschner, P. A. (2014). Learning ability development in flexible learning environments. In J. M. Spector, M. D. Merrill, J. Elen & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4 ed., pp. 363–372). New York: Routledge.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* (Vol. 534). Cham, Switzerland: Springer.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Cham, Switzerland: Springer.

Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., & Nardy, A. (2013). ReaderBench, an environment for analyzing text complexity and reading strategies. In H. C. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)* (pp. 379–388). Memphis, USA: Springer.

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L.K. (2017). Cohesion Network Analysis of CSCL Participation. *Behavior Research Methods*, 1–16. doi: 10.3758/s13428-017-0888-4

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015a). ReaderBench: An Integrated Cohesion-Centered Framework. In G. Conole, T. Klobucar, C. Rensing, J. Konert & É. Lavoué (Eds.), *10th European Conf. on Technology Enhanced Learning* (pp. 505–508). Toledo, Spain: Springer.

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015b). ReaderBench: The Learning Companion. In *17th Int. Conf. on Artificial Intelligence in Education (AIED 2015)* (pp. 915–916). Madrid, Spain: Springer.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M., & McNamara, D. S. (2015c). ReaderBench: An Integrated Tool Supporting both Individual and Collaborative Learning. In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* (pp. 436-437). Poughkeepsie, NY: ACM.

Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015a). Dialogism: A Framework for CSCL and a Signature of Collaboration. In O. Lindwall, P. Häkkinen, T.

Koschmann, P. Tchounikine & S. Ludvigsen (Eds.), *11th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2015)* (pp. 86–93). Gothenburg, Sweden: ISLS.

Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015b). Discourse cohesion: A signature of collaboration. In *5th Int. Learning Analytics & Knowledge Conf. (LAK'15)* (pp. 350–354). Poughkeepsie, NY: ACM.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423. doi: 10.1007/s11412-015-9226-y

Dessus, P., Gutu, G., Dascalu, M., Diouf, J. B., & Trausan-Matu, S. (2017). Vers des manuels de cours universitaires ouverts et interactifs promouvant l'apprentissage auto-régulé. In *Atelier "Evaluation formative pratiquée en classe ou en amphithéâtre" joint à l'ORPHEE*. Font-Romeu, France.

Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2003*. Chesapeake, VA, USA.

Ermoshina, K., Musiani, F., & Halpin, H. (2016, September). End-to-end encrypted messaging protocols: An overview. In *International Conference on Internet Science*. 244-254: Springer International Publishing.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1-32.

Garson, G. D. (2015). *Multivariate GLM, MANOVA, and MANCOVA*. Asheboro, NC: Statistical Associates Publishing.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik, 14*(5), 403–420.

Gutu, G., Dascalu, M., Rebedea, T., & Trausan-Matu, S. (2017). Time and Semantic Similarity – What is the Best Alternative to Capture Implicit Links in CSCL Conversations? In *12th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2017)* (pp. 223–230). Philadelphia, PA: ISLS.

Gutu, G., Dascalu, M., Ruseti, S., Rebedea, T., & Trausan-Matu, S. (2017). Unlocking the Power of Word2Vec for Identifying Implicit Links. In *17th IEEE Int. Conf. on Advanced Learning Technologies (ICALT2017)* (pp. 199–200). Timisoara, Romania: IEEE.

Gutu, G., Dascalu, M., Trausan-Matu, S., Heutelbeck, D., Hemmje, M., Westera, W., & Trausan-Matu, S. (2017). Semantic Annotation and Automated Text Categorization using Cohesion Network Analysis. In *5th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 13th Int. Conf. on eLearning and Software for Education (eLSE 2017)* (pp. 25–32). Bucharest, Romania: Advanced Distributed Learning Association.

Gutu, G., Rebedea, T., & Trausan-Matu, S. (2015). A comparison of semantic similarity techniques for a corpus of CSCL chats. In *RoEduNet International Conference-Networking in Education and Research (RoEduNet NER), 2015 14th* (pp. 178-183): IEEE.

Gutu, G., Ruseti, S., Dascalu, M., & Trausan-Matu, S. (2017). Keyword Mining and Clustering based on Cohesion Network Analysis. In D. Trandabat & D. Gifu (Eds.), *3rd Workshop on Social Media and the Web of Linked Data (RUMOUR 2017), in conjunction with the Joint Conference on Digital Libraries (JCLD 2017)* (pp. 23–30). Toronto, Canada: "Alexandru Ioan Cuza" University Publishing House.

Hanson, Stephen José, & Bauer, Malcolm. (1989). Conceptual Clustering, Categorization, and Polymorphy. *Machine Learning, 3*(4), 343-372. doi: 10.1023/a:1022697818275

Harris, Zellig S. (1954). Distributional Structure. *WORD, 10*(2-3), 146-162. doi: 10.1080/00437956.1954.11659520

Holmer, T., Kienle, A., & Wessner, M. (2006). Explicit Referencing in Learning Chats: Needs and Acceptance. In W. Nejdl & K. Tochtermann (Eds.), *Innovative Approaches for Learning and Knowledge Sharing, First European Conference on Technology Enhanced Learning, EC-TEL 2006* (pp. 170– 184). Crete, Greece: Springer.

John Walker, S. (2014). *Big data: A revolution that will transform how we live, work, and think*.

Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010). SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *5th Int. Workshop on Semantic Evaluation* (pp. 21–26): Association for Computational Linguistics.

Kinshuk. (2016). *Developing adaptive and personalized learning environments*. New York: NY: Routledge.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). Dirichlet and Inverted Dirichlet Distributions *Continuous Multivariate Distributions* (Vol. 1: Models and Applications, pp. 485–527). New York, NY: Wiley.

Kullback, S., & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Laskowski, P., & Chuang, J. (2006). Network monitors and contracting systems: competition and innovation. *ACM SIGCOMM Computer Communication Review, 36*(4), 183–194.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, MA: ACL.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). *Corpus-based and knowledge-based measures of text semantic similarity*. Paper presented at the 21st Int. Conf. AAAI, Boston, Massachusetts.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representation in Vector Space. In *Workshop at ICLR*. Scottsdale, AZ.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Nistor, N., Dascalu, M., Gutu, G., Trausan-Matu, S., Choi, S., Haberman-Lawson, A., Brands, B., Korner, C., & Koletzko, B. (in press). Mass Customization in Continuing Medical Education: Automated Extraction of E-Learning Topics. In *12th European Conference on Technology Enhanced Learning (EC-TEL 2017)*. Tallinn, Estonia: Springer.

Nistor, N., Dehne, A., & Drews, F. T. (2010). Mass customization of teaching and training in organizations. Design principles and prototype evaluation. *Studies in Continuing Education, 32*(3), 251-267.

Pine, B. J. (1993). *Mass customization: the new frontier in business competition*: Harvard Business Press.

Popescu, O., & Strapparava, C. (2015). SemEval 2015, Task 7: Diachronic Text Evaluation. In *9th Int. Workshop on Semantic Evaluation (SemEval 2015)* (pp. 870–878). Denver, Colorado: ACL.

Sagot, B. (2008). WordNet Libre du Francais (WOLF). Paris: INRIA. Retrieved from http://alpage.inria.fr/~sagot/wolf.html

Savický, P, & Hlavácová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*(3), 215-231. doi: 10.1076/jqul.9.3.215.14124

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Sebastiani, Fabrizio. (2002). Machine learning in automated text categorization. *ACM Comput. Surv., 34*(1), 1-47. doi: 10.1145/505282.505283

Soller, Amy. (2001). Supporting Social Interaction in an Intelligent Collaborative Learning System. *International Journal of Artificial Intelligence in Education (IJAIED), 12*, 40-62.

Stahl, Gerry. (2006). *Group cognition. Computer support for building collaborative knowledge*. Cambridge: MIT Press.

Swoboda, T., Hemmje, M., Dascalu, M., & Trausan-Matu, S. (2016). Combining Taxonomies using Word2vec. In *DocEng 2016* (pp. 131–134). Vienna, Austria: ACM.

Trausan-Matu, S., & Rebedea, T. (2010). A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants. In A. F. Gelbukh (Ed.), *11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010)* (pp. 354–363). New York: Springer.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, Indiana University Press, 6*(1), 58–74.

Trausan-Matu, S., Stahl, G., & Zemel, A. (2005). Polyphonic Inter-animation in Collaborative Problem Solving Chats. Philadelphia: Drexel University.

Winne, P. H. (2011). A cognitive and meta-cognitive analysis of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–32). New York: Routledge.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology, 3*(1), 1-191.

Winograd, T. (1980). What does it mean to understand language? *Cognitive science, 4*(3), 209-241.

Wu, X., Kumar, V., Quinlan, J. Ross, Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37.

Yarlett, Daniel G., & Ramscar, Michael J.A. (2008). Language Learning Through Similarity-Based Generalization. In.

Appendix I. Sample CSCL chat conversation

```
<Dialog team="34">
      <Body>
            <Turn nickname="Participant_1">
                  <Utterance genid="1" time="09.23.47" ref="0">joins the
room</Utterance>
            </Turn>
            <Turn nickname="Participant_2">
                  <Utterance genid="2" time="09.29.01" ref="0">joins the
room</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_3">
                  <Utterance genid="9" time="12.14.04" ref="0">hi
guys</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="10" time="12.14.18" ref="0">hello
everybody</Utterance>
            </Turn>
            ...
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="13" time="12.14.53" ref="0">so what is
out topic today boys?</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="14" time="12.15.03"
ref="0">collaborative learning right?</Utterance>
            </Turn>
            <Turn nickname="Participant_2">
                  <Utterance genid="15" time="12.15.12"
ref="0">yes</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="16" time="12.15.23" ref="0">so...which
are the actors...</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="17" time="12.15.32" ref="0">may i be
chat?:P</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="18" time="12.15.39" ref="0">i like
chatting</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="19" time="12.15.45"
ref="0">:D</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="20" time="12.15.51"
ref="0">you?</Utterance>
            </Turn>
            <Turn nickname="Participant_2">
                  <Utterance genid="21" time="12.16.01" ref="0">i`ll be
wiki</Utterance>
            </Turn>
```

```
            <Turn nickname="Participant_3">
                  <Utterance genid="22" time="12.16.07" ref="0">hi
wiki</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="23" time="12.16.38" ref="0">my name is
Participant_4 and i"ll be talking about blog</Utterance>
            </Turn>
            <Turn nickname="Participant_1">
                  <Utterance genid="24" time="12.17.02" ref="0">i'm
Participant_1 and i'll be forum</Utterance>
            </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="25" time="12.17.11" ref="0">hey
there</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_4">
                  <Utterance genid="27" time="12.17.26" ref="0">I'll go
first</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="28" time="12.17.32" ref="0">So, my
topic is blogging</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="29" time="12.17.42" ref="0">Blogs,
short for Web logs, are online writings that often invite reader comment
and criticism</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="30" time="12.17.52" ref="0">Postings
usually appear in reverse chronological order,</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_2">
                  <Utterance genid="96" time="12.38.51" ref="0">do forums
or blogs offer these facilities ?</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="97" time="12.39.00" ref="0">one thing
about wiki</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="98" time="12.39.15" ref="0">it is easy
for many people to post on them</Utterance>
            </Turn>
            <Turn nickname="Participant_1">
                  <Utterance genid="99" time="12.39.31" ref="96">you can
also find good information on forums</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_2">
                  <Utterance genid="103" time="12.41.05" ref="102">in
forums ,knowledge tends to be dispersed and somewhat lower in
density</Utterance>
            </Turn>
            <Turn nickname="Participant_1">
```

```
                  <Utterance genid="104" time="12.41.14" ref="103">i
agree</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_3">
                  <Utterance genid="119" time="12.45.27" ref="118">as
opposed to forum..where yout answer will hopefully come</Utterance>
            </Turn>
            <Turn nickname="Participant_1">
                  <Utterance genid="120" time="12.45.40" ref="119">but on
wiki you can't ask questions about a certain matter</Utterance>
            </Turn>
            <Turn nickname="Participant_2">
                  <Utterance genid="121" time="12.46.03" ref="117">but for
chat, users have to be online at the same time.</Utterance>
            </Turn>
            ...         </Turn>
            <Turn nickname="Participant_3">
                  <Utterance genid="183" time="13.00.42" ref="0">all
technologies combined in one framework</Utterance>
            </Turn>
            <Turn nickname="Participant_2">
                  <Utterance genid="184" time="13.00.57" ref="183">it could
be a great idea an integrated environment</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_3">
                  <Utterance genid="186" time="13.13.13" ref="0">yup...i
agree</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_4">
                  <Utterance genid="188" time="13.13.38" ref="0">i can say,
that all of them, blog, chat, forum and wiki have advantages and
disadvantages</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="189" time="13.02.00" ref="0">but
togheter i think that it will be a flawless system, or almost an flawless
system</Utterance>
            </Turn>
            ...
            <Turn nickname="Participant_3">
                  <Utterance genid="195" time="13.02.58" ref="0">It was a
very nice chat, I enjoyed, we'll talk later</Utterance>
            </Turn>
            <Turn nickname="Participant_4">
                  <Utterance genid="196" time="13.03.04" ref="0">goodbye
everybody</Utterance>
            ...
            <Turn nickname="Participant_2">
                  <Utterance genid="203" time="13.03.33" ref="0">leaves the
room</Utterance>
            </Turn>
      </Body>
</Dialog>
```