Chapter 4
Computational Discourse Analysis

As previous chapters were overall oriented towards comprehension and productions from the perspectives of individual and collaborative learning, this chapter is focused on presenting automatic discourse analysis models and natural language processing techniques that ground a computational and quantifiable perspective of cohesion and coherence and that greatly impact the underlying functionalities of our developed systems (*A.S.A.P.*, *Ch.A.M.P.*, *PolyCAFe* and *ReaderBench*).

4.1 Measures of Cohesion and Local Coherence

From a computational viewpoint, we limit the perspective of coherence and cohesion (see *2.1.1* Coherence and Cohesion) to *lexical and semantic cohesion* and *local coherence* that captures text organization at the level of sentence to sentence transitions, further necessary to achieve global coherence (Lapata and Barzilay 2005). In this computational context, *cohesion* is reflected in the linguistic form of discourse (McNamara et al. 2010) and is often regarded as an indicator of its structure. More specifically, cohesion can derive from: 1/ discourse connectedness through cue words or phrases (e.g., "but", "because") as relations between sentences (e.g., explanation, contrast); 2/ referencing expressions that reflect the status of an entity in the discourse and can be identified through co-reference resolution (Jurafsky and Martin 2009; Raghunathan et al. 2010); 3/ lexically or semantically related words obtained from semantic distances in ontologies (Budanitsky and Hirst 2006) (see *4.3.1* Semantic Distances and Lexical Chains), cosine similarity in vector spaces from Latent Semantic Analysis (Landauer et al. 1998a) (see *4.3.2* Semantic Similarity through Tagged LSA) or through topic relatedness in Latent Dirichlet Allocation (Blei et al. 2003) (see *4.3.3* Topic Relatedness through Latent Dirichlet Allocation). Aligned with the previous definition are also the two measures of textual cohesion proposed by Graesser et al. (2004), frequently used in automated discourse analysis: *referential cohesion* (the degree to which words, concepts or phrases are related or repeated across the text) and *causal cohesion* (marked by the explicit use of connectives – e.g., "since", "because", "therefore", "the cause of" or "as a consequence").

Coherence, on the other hand, is much more difficult to express from a computational perspective as multiple levels that simultaneously relate discourse elements need to be taken into consideration (Grosz and Sidner 1986). Moore and

M. Dascalu, *Analyzing Discourse and Text Complexity for Learning and Collaborating*,
Studies in Computational Intelligence 534,
DOI: 10.1007/978-3-319-03419-5_4, © Springer International Publishing Switzerland 2014

Pollack (1992) focus on two levels in particular: 1/ the *informational level*, mostly centered on causal relations between utterances, weakly related to the linguistic form and difficult to model in comparison to previous links between words; and 2/ the *intentional level*, aimed at the changes in the discourse participants' mental states, superficially visible in the linguistic form and extremely difficult to model in terms of computational analysis. Moreover, the same study highlights also a problem of the rhetorical structure theory (Mann and Thompson 1987) (see *4.2* Discourse Analysis and the Polyphonic Model) that is limited to a single, preferred rhetorical relation between consecutive discourse elements, whereas coherence should be modeled as an overlap of multiple relations between the same text spans, but at different levels. Nevertheless, while addressing the informational level, coherence is most frequently accounted by: lexical chains (Morris and Hirst 1991; Barzilay and Elhadad 1997; Lapata and Barzilay 2005) (see *4.3.1* Semantic Distances and Lexical Chains), centering theory (Miltsakaki and Kukich 2000; Grosz et al. 1995) (see *4.2* Discourse Analysis and the Polyphonic Model) in which coherence is established via center continuation, or Latent Semantic Analysis (Foltz et al. 1993, 1998) (see *4.3.2* Semantic Similarity through Tagged LSA) used for measuring the cosine similarity between adjacent phrases; in the end, overall coherence is considered the mean value of the previous semantic similarities. Nevertheless, from a computational perspective and through its intrinsic nature consisting of a bag-of-words approach, LSA fundamentally supports cohesion and not coherence.

4.2 Discourse Analysis and the Polyphonic Model

Discourse may be defined as "a coherent structured group of sentences" (Jurafsky and Martin 2009, ch. 21) that in NLP is usually considered different in monologues and dialogues. However, in both cases the same idea of an emitter–receiver channel is used, the difference being the uni-respectively bi-directional communications (Trausan-Matu and Rebedea 2010). Therefore, one-way, speaker-listener directed models of communication are considered in monologues (Jurafsky and Martin 2009). The usual way of analyzing discourse in this case is the segmentation of text, the search for different relationships among segments, the measurement of coherence and obtaining some discourse abstractions like co-references or summaries. In this context, cohesion seen as lexical, grammatical and semantic links between textual fragments becomes a central element, whereas coherence is considered as granted, in different degrees, when analyzing texts. On the other hand, the detection of local relations can be used for measuring coherence. Some structures are searched, as the Rhetorical Structure Theory (RST) (Mann and Thompson 1987, 1988), which considers a hierarchical decomposition of a text. Centering Theory (Grosz et al. 1995) and co-reference resolution systems (Jurafsky and Martin 2009) may be also considered (Trausan-Matu and Rebedea 2010).

On the other hand, dialogue analysis has as prototype phone-like (or face-to-face) conversations. A typical approach starts from analyzing local, two-participant data and tries to identify speech acts, dialog acts and afterwards, adjacency pairs (Jurafsky and Martin 2009). Even if there are attempts to analyze conversations with multiple participants, considering a more global, collaboration-based perspective, like transacts (Joshi and Rosé 2007), the approach is also based on a two interlocutors' model (Trausan-Matu and Rebedea 2010).

In terms of discourse analysis, probably the most known discourse theories belong to Hobbs (1985), Grosz et al. (1995) or Mann and Thompson (1987). Hobbs' theory is based on considering semantic coherence relations – "a set of binary relations between a current utterance and the preceding discourse" (Hobbs 1978) – and on using abduction inferences in formal logic (Hobbs 1985, 1979). "Coherence thus plays a role beyond sentence boundaries analogous to the role played by grammaticality within sentences. It is the mortar with which extended discourse is constructed." (Hobbs 1979). Also, of particular interest is the phenomenon of topic drifting observed in spoken conversations – although adjacent segments are coherent, the end of the conversation is significantly different from the starting point – that is mainly induced by three mechanisms: sematic parallelism, chained explanations and metatalk (Hobbs 1990).

Rhetorical Structure Theory (RST) (Mann and Thompson 1987) identifies hierarchical rhetorical structures between text spans (defined as any contiguous interval of text), classified as nuclei or satellites in accordance to their importance, that is built using a limited set of rhetorical schemas (patterns) like antithesis and concession, elaboration, enablement and motivation, interpretation and evaluation, restatement and summary, etc. The theory requires the fulfillment of 4 constraints for a successful RST analysis (Mann and Thompson 1987): 1/ completeness as coverage of the entire text, 2/ connectedness focusing on the recursive division of text spans, 3/ uniqueness as each relation is applied on different text spans and 4/ adjacency as adjoined text spans are consecutive.

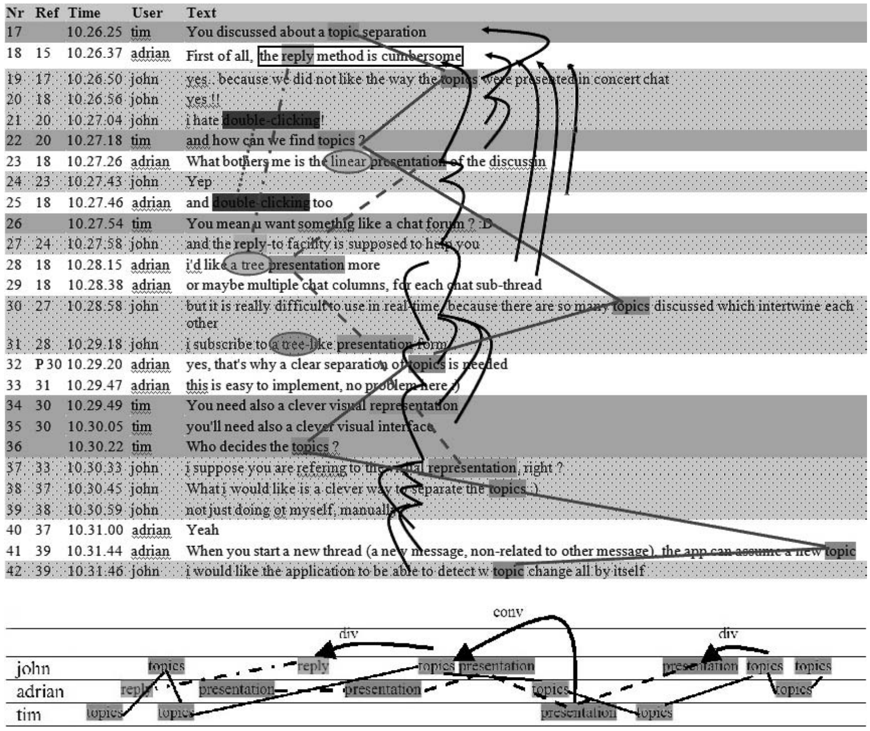
From a different perspective, coherence is obtained in the centering theory (Grosz et al. 1995) at both local (coherence among the utterances in a given segment) and global levels (coherence with other segments of the discourse), centered on two different aspects: intentional and attentional states, which together with the linguistic structure of an utterance sequence, form a tripartite organization. An intentional structure should be present in each discourse, assuring that discourse is rational. This structure is built from intentions (purposes) and, sometimes, from the beliefs of the author of the discourse (or of each participant in a conversation) and from relationships among linguistic segments (Grosz et al. 1995). In addition, two types of centers are identified: backward-looking and forward-looking. Continuation of the discourse is modeled through an ordered set of forward-looking centers defined at utterance level plus a single back-looking center (except for the first utterance of the discourse segment), "that provides a coherent link to the previous utterances by being coreferential with one of the forward-looking centers of that utterance" (Gordon et al. 1993).

On the other hand, the polyphonic theory (Trausan-Matu et al. 2005; Trausan-Matu 2010c; Trausan-Matu and Stahl 2007; Trausan-Matu and Rebedea 2009; Trausan-Matu et al. 2010b) follows the ideas of Koschmann (1999) and Wegerif (2005) and investigates how Bakhtin's theory of polyphony and inter-animation (Bakhtin 1981, 1984) (see *3.1.2* Bakhtin's Dialogism) can be used for analyzing the discourse in chat conversations with multiple participants. In phone and face-to-face dialogs only one person usually speaks at a given moment in time, generating a single thread of discussion. This is, of course, determined by the physical, acoustical constraints (if two or more persons are speaking in the same moment, it is impossible to understand something). In chat environments, like the one used in the Virtual Math Teams (VMT) project (Stahl 2009a), any number of participants may write utterances at the same time. As discussed in a previous section, the VMT environment offers also explicit referencing facilities that allow the users to indicate to what previous utterance(s) they refer to (see *3.1.1* Chats as Support for Social Cognition). This facility is extremely important in chat conversations with more than two participants because it allows the existence of several discussion threads in parallel. Moreover, the co-occurrence of several threads gives birth to inter-animation, a phenomenon similar to polyphony, where several voices jointly play a coherent piece as a whole (Trausan-Matu et al. 2007a; Trausan-Matu and Rebedea 2009).

Bakhtin (1984) emphasized that polyphony occurs in any text. He considered that dialog characterizes any text, that "our speech, that is, all our utterances (including creative works), is filled with others' words" (Bakhtin 1986). The voice becomes a central concept, has a more complex meaning. A voice is not limited to the acoustic dimension, it may be considered as a particular position, which may be taken by one or more persons when emitting an utterance, which may have both *explicit*, similar to those provided by the VMT chat environment (Stahl 2009a), and *implicit* links (for example, lexical chains, co-references or argumentation links) and influence other voices. Each utterance is filled with 'overtones' of other utterances (Stahl 2009a). Moreover, by the simple fact that they co-occur, voices are permanently inter-animating, entering in competition, generating multi-vocality in any conversation and even in any text (in Bakhtin's dialogic theory everything is a dialog) or, as Bakhtin calls it, a "heteroglossia, which grows as long as language is alive" (Bakhtin 1981).

The ideas of Bakhtin drive to a musical metaphor for discourse and for learning: "the voices of others become woven into what we say, write, and think" (Koschmann 1999). Therefore, for analyzing discourse in chats the aim shifts towards investigating how voices are woven, how themes and voices inter-animate in a polyphonic way (Trausan-Matu et al. 2007b). This is important not only for understanding how meaning is created but also for trying to design tools for support and evaluation. Figure 7 presents the inter-animation of voices within a chat conversation and their evolution in time, following a pattern first described by Trausan-Matu et al. (2005); the longest two voices are represented by the linked curly lines. As it can be observed, several threads can co-appear in parallel and even the same participant may participate to more than one discussion thread within a given timeframe (e.g. John, at utterance 19, approves and elaborates Tim's intervention, while in the following utterance represents an approval of Adrian's utterance 18) (Trausan-Matu 2010c). Therefore, this co-presence of multiple discussion threads and their inter-influences models voice inter-animation towards achieving polyphony.

The polyphonic model focuses on the idea of identifying voices in the analysis of discourse and building an internal graph-based representation, whether we are focusing on the utterance graph (Trausan-Matu et al. 2007a) or the cohesion graph (Trausan-Matu et al. 2012a; Dascalu et al. 2013a) (see *7.2* Cohesion-based Discourse Analysis). For this aim, links between utterances are analyzed using adjacency pairs, repetitions, lexical chains, speech and argumentation acts or cohesive links, a graph is built from which discussion threads are identified. Nevertheless, in both internal representations, lexical or semantic cohesion between any two utterances seen as explicit communicative acts can be considered the central liaison between the analysis elements within the graph. Cohesion can be expressed as the "distance" between the utterance boundaries (Dong 2005) and can be computed by various means of semantic similarity, including semantic distances in ontologies (see *4.3.1* Semantic Distances and Lexical Chains), latent vector space representations (see *4.3.2* Semantic Similarity through Tagged LSA) or topic models (see *4.3.3* Topic Relatedness through Latent Dirichlet Allocation).



Nr	Ref	Time	User	Text
17		10.26.25	tim	You discussed about a topic separation
18	15	10.26.37	adrian	First of all, the reply method is cumbersome
19	17	10.26.50	john	yes.. because we did not like the way the topics were presented in concert chat
20	18	10.26.56	john	yes !!
21	20	10.27.04	john	i hate double-clicking!
22	20	10.27.18	tim	and how can we find topics ?
23	18	10.27.26	adrian	What bothers me is the linear presentation of the discussin
24	23	10.27.43	john	Yep
25	18	10.27.46	adrian	and double-clicking too
26		10.27.54	tim	You mean u want somethig like a chat forum ? :D
27	24	10.27.58	john	and the reply-to facility is supposed to help you
28	18	10.28.15	adrian	i'd like a tree presentation more
29	18	10.28.38	adrian	or maybe multiple chat columns, for each chat sub-thread
30	27	10.28.58	john	but it is really difficult to use in real-time, because there are so many topics discussed which intertwine each other
31	28	10.29.18	john	i subscribe to a tree-like presentation form
32	P 30	10.29.20	adrian	yes, that's why a clear separation of topics is needed
33	31	10.29.47	adrian	this is easy to implement, no problem here :)
34	30	10.29.49	tim	You need also a clever visual representation
35	30	10.30.05	tim	you'll need also a clever visual interface
36		10.30.22	tim	Who decides the topics ?
37	33	10.30.33	john	i suppose you are refering to the visual representation, right ?
38	37	10.30.45	john	What i would like is a clever way to separate the topics :)
39	38	10.30.59	john	not just doing ot myself, manually
40	37	10.31.00	adrian	Yeah
41	39	10.31.44	adrian	When you start a new thread (a new message, non-related to other message), the app can assume a new topic
42	39	10.31.46	john	i would like the application to be able to detect w topic change all by itself

Fig. 7 Inter-animation of voices within a chat (Trausan-Matu et al. 2007b)

As the initial polyphonic model used the utterance graph (Trausan-Matu et al. 2007a) and the cohesion graph (Trausan-Matu et al. 2012a; Dascalu et al. 2013a), which can be seen as a generalization, is presented in detail in *7.2* Cohesion-based Discourse Analysis, we will focus on providing a comprehensive view of the polyphonic model, using as underlying representation the utterance graph. This internal structure is built upon two types of links between utterances: explicit and implicit. Participants add manually explicit links during their chat sessions by using a facility from the conversation environment – e.g., Concert Chat (Holmer et al. 2006). On the other hand, implicit links are automatically identified by means of co-references, repetitions, lexical chains and inter-animation patterns (Trausan-Matu et al. 2005; Trausan-Matu and Rebedea 2010). In the resulted graph, each utterance is a node and the weights of edges are given by the similarity between the utterances. The orientation of each edge follows the timeline of the chat and the evolution of the discussion in time. Starting from the previous graph, a thread can be easily identified as a logical succession of explicitly or implicitly inter-linked utterances. Moreover, the primary extension of each utterance is its inner voice that inter-twines with other voices from the same thread or from different ones, but with less strength. A new intervention or a new utterance in terms of units of analysis can be clearly expressed as a voice and aspects that need to be addressed include: degree of interconnection in terms of cohesion with other utterances, relevance within the discourse or future impact in the overall discussion.

Starting from Bakhtin (1984) perspective of discourse analysis, each identified voice may become more or less powerful than the others and may influence the others. Among chat voices there are sequential and transversal relations, highlighting a specific point of view in a counterpointal way, as mentioned in previous work (Trausan-Matu and Rebedea 2009). The cooccurrence of several voices which enter in dialogue is a phenomenon considered by Bakhtin to be universal, present in any text, not only in conversations: "Life by its very nature is dialogic … when dialogue ends, everything ends" (Bakhtin 1984). Bakhtin moves the focus of analysis from sentences to utterances in an extended way, in which even an essay contains utterances and is, at its turn, an utterance. Moreover, each utterance is filled with 'overtones' that contain the echoes and influence of other previous utterances.

A voice is generated by an utterance with effects (echoes) on the subsequent utterances via explicit and implicit links. Moreover, by the simple fact that they co-occur, voices are permanently interacting, overlapping and inter-animating, entering in competition, and generating multivocality in any conversation. The ideal situation of a successful conversation or a coherent discourse is achieved when the voices are entering inter-animation patterns based on the discussion threads they are part of (Trausan-Matu et al. 2005).

Moreover, of particular interest is the multi-dimensionality of the polyphonic model (Trausan-Matu 2013). Firstly, the *longitudinal* dimension is reflected in the explicit or implicit references between utterances, following the conversation timeline. This grants an overall image of the degree of inter-animation of voices spanning the discourse, which can later on be particularized as collaboration, seen as the interactions between multiple participants of the conversation reflected in their voices. Secondly, *threading* highlights voices evolution in terms of the interaction with other discussion threads. Thirdly, the *transversal* dimension is useful for observing a differential positioning of participants, when a shift of their point of interest occurs towards discussing other topics. In the end, this combination of continuity (longitudinal dimension) versus juxtaposition (transversal dimension) of voices, respectively centrifugal versus centripetal forces exerted by participants in terms of covered concepts generates polyphony.

In addition, the co-presence of multiple voices in the same time inherently generates consonances and dissonances, similarly to the polyphonic musical case. In this context, these inter-animation effects of consonance and dissonance in voices overlap can be perceived as centripetal and centrifugal forces tightly correlated in the trend of achieving discourse coherence. The weaving of the voices all along the longitudinal time dimension and meanwhile their consonance/ dissonance on the transversal dimension is similar to the case of polyphonic music (Trausan-Matu et al. 2006): ”The deconstructivist attack […] – according to which only the difference between difference and unity […] can act as the basis of a differential theory […] – is the methodical point of departure for the distinction between polyphony and non-polyphony.” (Mahnkopf 2002)

From a computational perspective, until recently, the goals of discourse analysis in existing approaches oriented towards conversations analysis were to detect topics and links (Adams and Martell 2008), dialog acts (Kontostathis et al. 2009), lexical chains (Dong 2006) or other complex relations (Rosé et al. 2008) (see *3.1.3* CSCL Computational Approaches). The polyphonic model takes full advantage of term frequency – inverse document frequency *Tf-Idf* (Adams and Martell 2008; Schmidt and Stone), Latent Semantic Analysis (Schmidt and Stone ; Dong 2006), Social Network Analysis (Dong 2006), Machine Learning (e.g., Naïve Bayes (Kontostathis et al. 2009), Support Vector Machines and Collin's perceptron (Joshi and Rosé 2007), the *TagHelper* environment (Rosé et al. 2008) and the semantic distances from the lexicalized ontology *WordNet* (Adams and Martell 2008; Dong 2006). The model starts from identifying words and patterns in utterances that are indicators of cohesion among them and, afterwards, performs an analysis based on the graph, similar in some extent to a social network, and on threads and their interactions.

As conclusion, the polyphonic discourse analysis model, built on Bakhtin's dialogism and supported by multiple natural language processing techniques (presented in detail in *4.3* Natural Language Processing Techniques) can be considered a viable representation of discourse, with emphasis on the analysis of multi-participant conversations for which classic approaches are not appropriate. Moreover, initial validations performed by Trausan-Matu (2011) showed that the results of the polyphonic analysis were close to those of tutors, whereas its extension in terms of assessing collaboration (see *9.2* Collaboration Assessment) proves its applicability.

4.3 Natural Language Processing Techniques

While addressing natural language processing techniques (Manning and Schütze 1999), of particular interest is to what extent computational models of semantic memory (Cree and Armstrong 2012) grasp underlying semantic relations and meanings of concepts from texts, and how these models can be effectively used to measure cohesion between textual fragments (Bestgen 2012). In this context, three complementary approaches are most remarkable: 1/ semantic distances in ontologies (Budanitsky and Hirst 2006), 2/ semantic vector spaces extracted through Latent Semantic Analysis (LSA) (Landauer and Dumais 1997) and 3/ probabilistic topics modeling by using Latent Dirichlet Allocation (Blei et al. 2003), presented in detail in the current section and integrated in various developed systems. The presentation of each approach offers a broad perspective of the method and of the used resources, particularities, possible improvements and drawbacks.

4.3.1 Semantic Distances and Lexical Chains

As knowledge can be formally represented as a conceptualization consisting of objects, concepts or other entities presumably related to an area of interest and of relationships linking them together (Genesereth and Nilsson 1987), an ontology can be seen as an "explicit specification of a conceptualization" (Gruber 1993). Therefore, an ontology consists of a set of concepts specific to a domain and of the relations between pairs of concepts. Starting from the representation of a domain, we can define various distance metrics between concepts based on the defined relationships among them and later on extract lexical chains, specific to a given text that consist of related/cohesive concepts spanning throughout a text fragment or the entire document.

A Lexicalized Ontologies and Semantic Distances

One of the most commonly used resources for English sense relations in terms of lexicalized ontologies is the *WordNet* lexical database (Fellbaum 1998; Miller 1995, 2010) that consists of three separate databases, one for nouns, a different one for verbs, and a third one for adjectives and adverbs. *WordNet* groups words into sets of cognitively related words (synsets), thus describing a network of meaningfully inter-linked words and concepts. Therefore, synonymy is the main relation between words that are now grouped into unordered sets that also include a brief description or gloss, useful for word sense disambiguation (WSD) (Navigli 2009).

In addition, *WordNet* is built using the principle of "cognitive plausibility" as the organization of words mimics cognitively related concepts (Miller 1998; Emond 2006). This principle of plausibility is based on three hypotheses: *separability* – "lexical knowledge is independent from other language related knowledge"; *patterning* – "relations and patterns between lexical entities are central to natural language processing" and *comprehensiveness* – "any computation model of human language processing should have a store of lexical knowledge as extensive as people do" (Miller 1998; Emond 2006).

Synsets are interconnected using semantic relations that vary based on the underlying part-ofspeech (see Figure 8 and Table 4). In addition, the internal organization of nouns and verbs uses a hierarchy built on ”IS A” relationships and the links between synsets can be regarded as specialization relations between conceptual categories, aligning the perspectives of *WordNet:* lexical database versus lexicalized ontology. As an overview of *WordNet*, each database consists of a set of lemmas annotated with a set of corresponding senses, covering in the 3.0 version approximately 117k nouns, 11k verbs, 22k adjectives and 5k adverbs; the average noun has 1.23 senses, while verbs have 2.16 senses on average.

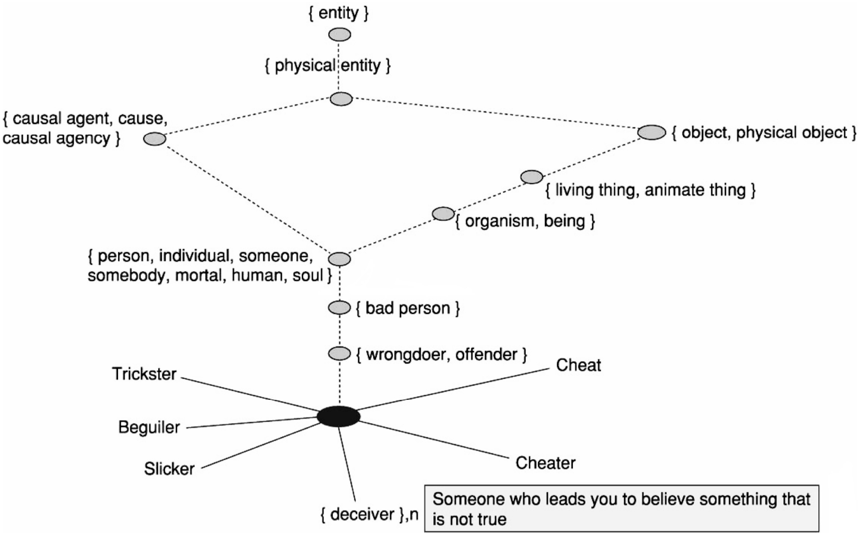

{ entity }
{ physical entity }
{ causal agent, cause, causal agency }
{ object, physical object }
{ living thing, animate thing }
{ organism, being }
{ person, individual, someone, somebody, mortal, human, soul }
{ bad person }
{ wrongdoer, offender }
Trickster
Cheat
Beguiler
Slicker
Cheater
{ deceiver },n
Someone who leads you to believe something that is not true

Fig. 8 *WordNet* noun tree reflecting semantic/hierarchical relations (Fellbaum 2005, p. 666)

Table 4 Word part-of-speech and relations between synsets in *WordNet* (Fellbaum 2005)

Word part-of-speech	Available relations between synsets
Noun	*hypernymy* – "is a" generalization *hyponymy* – "is a" specialization *coordination/sibling* – concepts share a hypernym *holonymy* – "is a part of" generalization *meronymy* – "is a part of" specialization
Verb	*Entailment relationships* *troponymy*- one activity expresses a particular manner of the other *backward entailment, presupposition* and *cause*
Adjective	*Descriptive adjective* *direct antonymy* and *indirect antonymy* *Relational adjective* *related noun*
Adverb	*base adjective*

Regarding other freely available similar resources, *WordNet Libre du Francais – WOLF* (Sagot 2008; Sagot and Darja 2008) is the best French alternative that uses the XML file format developed within the IST-2000-29388 BalkaNet – Design and Development of a Multilingual Balkan WordNet project (http://www.dblab.upatras.gr/balkanet/).

Besides word sense disambiguation, *WordNet* or similar resources are useful for determining the relatedness between concepts through semantic distances (Budanitsky and Hirst 2001, 2006; Wang and Hirst 2011; Pedersen et al. 2004) (see Table 5), query expansion using lexical-semantic relations (Voorhees 1994; Navigli and Velardi 2003; Moldovan and Mihalcea 2000) or the identification of speech acts (Yeh et al. 2008; Trausan-Matu and Rebedea 2010). Although multiple semantic distances exist and more can be added to the list presented in Table 5, there is no clear measure that best fits all analysis scenarios as "lexical semantic relatedness is sometimes *constructed* in context and cannot always be determined purely from an a priori lexical resource such as *WordNet*" (Murphy 2003; Budanitsky and Hirst 2006).

Nevertheless, we must also present the limitations of *WordNet* and of semantic distances, with impact on the development of subsequent systems (see *6 PolyCAFe* – Polyphonic Conversation Analysis and Feedback and *7 ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism): 1/ the focus only on common words, without covering any special domain vocabularies; 2/ reduced extensibility as the serialized model makes difficult the addition of new domain-specific concepts or relationships; 3/ most relations are between words with the same corresponding part-of-speech, significantly reducing the horizon for comparing the semantic relatedness between concepts; 4/ semantic problems or limitations, specific to a given context, that require additional cleaning – the *OntoClean* approach (Oltramari et al. 2002) and 5/ the encoded word senses are too fine-grained even for humans to distinguish different valences of particular concept senses, reducing the performance of WSD systems. For the later granularity issue, multiple clustering methods that automatically group together similar senses of the same word have been proposed (Agirre and Lopez 2003; Navigli 2006; Snow et al. 2007). In addition, when considering *WOLF* in which glosses are only partially translated, integrating in the end a mixture of both French and English definitions, only a limited number of semantic distances are applicable (e.g., path length, Leacock-Chodorow's normalized path length or Wu-Palmer as the most representative).

B Building the Disambiguation Graph

Lexical chaining derives from textual cohesion (Halliday and Hasan 1976) and involves the selection of related lexical items in a given text (e.g., starting from Figure 8, the following lexical chain could be generated if all words occur in the initial text fragment: "cheater, person, cause, cheat, deceiver, …"). In other words, the lexical cohesive structure of a text can be represented as lexical chaining that consists of sequences of words tied together by semantic relationships and that can span across the entire text or a subsection of it. The identified lexical chains are

Table 5 Semantic distances applied on *WordNet*

Name and reference	Formula	Description
Path length	$l(c_1, c_2)$	The length of the shortest path between two concepts/synsets.
Depth	$d(c_1) = l(c_1, root)$	The length of the path from the current concept to the global root.
Hirst-St-Onge (Hirst and St-Onge 1997)	$rel_{HS}(c_1, c_2) = C - l(c_1, c_2) - k \times dir$	Two words are considered semantically related if the path is not too long and its direction does not change too often (*dir* – number of direction changes; *k*, *C* – constants).
Leacock-Chodorow (Leacock and Chodorow 1998)	$sim_{LC}(c_1, c_2) = -log \frac{l(c_1, c_2)}{2D}$	The path length is normalized by the overall depth *D* of the ontology.
Resnik (Resnik 1995)	$sim_R(c_1, c_2) = -log\ (p(lso(c_1, c_2))$	Similarity is expressed as the information content of their lowest super-ordinate (*lso(c₁,c₂)* – most specific common sub-summer; *p(c)* – probability of occurrence of synset *c* in a specific corpus).
Jiang-Conrath (Jiang and Conrath 1997)	$dist_{JC}(c_1, c_2) = 2 \times log\left(p(lso(c_1, c_2))\right) - log(p(c_1)) - log(p(c_2))$	Besides the consideration of the most specific sub-summer, the information content of the two nodes also plays an important role in estimating the inverse of similarity.
Lin (Lin 1998)	$sim_L(c_1, c_2) = \frac{2 \times log\left(p(lso(c_1, c_2))\right)}{\log(p(c_1)) + log(p(c_2))}$	The measure follows the idea of similarity between objects, combined with $dist_{JC}$.
Wu-Palmer (Wu and Palmer 1994)	$sim_{WP}(c_1, c_2) = \frac{2 \times d(lso(c_1, c_2))}{l(c_1, lso(c_1, c_2)) + l(c_2, lso(c_1, c_2)) + 2 \times d(lso(c_1, c_2))}$	Conceptual similarity is a scaled metric perceived in comparison to a global depth.
Lesk (Banerjee and Pedersen 2002)	$sim_{Lesk}(c_1, c_2)$	Similarity is determined as an adaptation of the Lesk (1986) approach to WordNet by using the overlap between concept descriptions or glosses.

independent of the grammatical structure of the initial text and, in effect, the contained concepts from each chain capture a portion of the cohesive structure of the text. A lexical chain can provide a context for the resolution of an ambiguous term and enable identification of the concept that the term represents. In a particular manner, the lexical cohesive relationships between words can be established using a lexicalized ontology – *WordNet* (Miller 1995; Fellbaum 1998) or *WordNet Libre du*

Francais – WOLF (Sagot 2008). Since first proposed (Morris and Hirst 1991), lexical chains have been used in a variety of applications in the fields of Information Retrieval (IR) (Manning et al. 2008) and Natural Language Processing (Manning and Schütze 1999), most notably for word disambiguation (Galley and McKeown 2003), detection of malapropisms (Hirst and St-Onge 1997) and text summarization (Barzilay and Elhadad 1997; Silber and McCoy 2003).

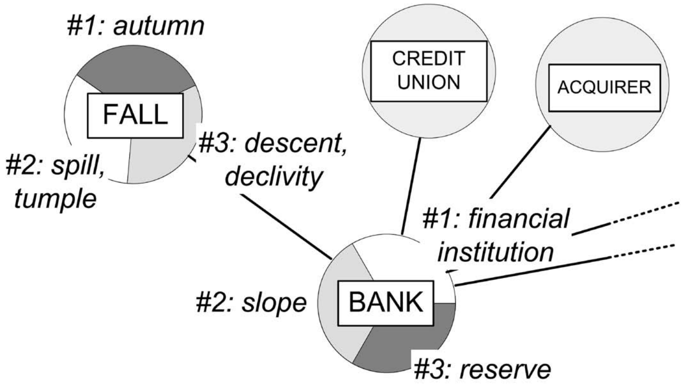


Fig. 9 Disambiguation graph example (Galley and McKeown 2003, p. 1487). Highlighting a possible implicit representation of word-sense combinations (#*n* denotes a word-sense)with all edge weights equal to 1.

Once the pre-processing of a given text is completed (splitting, tokenizing, part of speech tagging, parsing, named entity recognition, co-reference resolution) (Manning and Schütze 1999), the disambiguation graph (see Figure 9) can be built in linear time (Galley and McKeown 2003). In this kind of graph, nodes represent word instances and weighted edges represent semantic relations. Since *WordNet* or *WOLF* do not relate words but senses, each node is split into as many senses as the concept has, and each edge connects exactly two senses. In essence, if a word has *n* possible senses, it will initially have *n* different lexical chain links associated with it. Afterwards, when adding a new lexical chain link to the disambiguation graph, new connections need to be added between the concept and all the other related links in the graph.

The types of semantic relations taken into consideration when linking two words are hypernymy, hyponymy, synonymy, antonymy, or whether the words are siblings by sharing a common hypernym. The weights associated with each relation vary according to the strength of the relation and the proximity of the two words in the text analyzed. Table 6 depicts the weights later used in *ReaderBench* (see *7 ReaderBench* (1) – Cohesion-based Discourse Analysis and Dialogism), similar to Galley and McKeown (2003), but with antonymy having importance (and associated weights) equivalent to the synonymy relation.

Table 6 Lexical chains – adapted weights based on semantic relations and word distances (after Galley and McKeown 2003)

	Distance between words			
Semantic relations	1 sentence	3 sentences	same block/paragraph	other
Synonym/Antonym	1	1	.5	.5
Hypernym/Hyponym	1	.5	.3	.3
Sibling	1	.3	.2	0

The pruning of the disambiguation graph corresponds to the actual disambiguation step of the algorithm (Galley and McKeown 2003). Therefore, for each word, the values of the lexical chain links associated with each of the word senses are compared and the link with the best value is selected. The value of a link is computed as the sum of the weights of all the connections for that link or, in terms of the generated graph, the sum of the weights of all the edges connecting that link to other links in the graph. In the end, when a specific word is associated to a link, it has been disambiguated. The last step consists of removing all other links associated with the word's other senses, from the disambiguation graph. In order to optimize the process of identifying the link with the best value, these values can be computed incrementally when building the disambiguation graph, as new connection between two links are added. In this particular context, each lexical chain is, in fact, in itself a graph or, to be more exact, a connected component of the pruned disambiguation graph. Therefore, lexical chains are identified as connected components within the disambiguation graph by using the breadth-first search algorithm (Cormen et al. 2009).

4.3.2 Semantic Similarity through Tagged LSA

Latent Semantic Analysis (LSA) (Deerwester et al. 1989; Deerwester et al. 1990; Dumais 2004; Landauer and Dumais 1997) is a natural language processing technique starting from a vector-space representation of semantics highlighting the co-occurrence relations between terms and containing documents, after that projecting the terms in sets of concepts (semantic spaces) related to the initial texts. LSA builds the vector-space model, later on used also for evaluating similarity between terms and documents, now indirectly linked through concepts (Landauer et al. 1998a; Manning and Schütze 1999). Moreover, LSA can be considered a mathematical method for representing words' and passages' meaning by analyzing in an unsupervised manner a representative corpus of natural language texts. More formally, LSA uses a sparse term-document matrix that describes the occurrence of terms in corresponding documents. LSA performs a "bag-of-words" approach as it disregards word order by counting only term occurrences, later to be normalized. The indirect link induced between groups of terms and documents is obtained through a singular-value decomposition (SVD) (Golub and Kahan 1965; Golub and

Reinsch 1970; Landauer et al. 1998b) of the matrix, followed by a reduction of its dimensionality by applying a projection over *k* predefined dimensions, similar to the least-squares method (see Figure 10).

From a cognitive point of view, LSA has been thoroughly analyzed, with two prominent directions. Firstly, LSA can be seen as an expression of meaning as each word can be represented as a context-free vector in the semantic vector-space model (Kintsch 2000, 2001). The actual dimensions of concepts do not bear a specific individual meaning, but the overall representation generated by LSA can be considered a map of meanings (Landauer et al. 2007). Secondly, the semantic proximity effect (Howard and Kahana 1999) highlights the positive correlation between the similarities measured through LSA and the human recall using word association lists. Moreover, it was noted that the inter-response time between similar words was much quicker than for dissimilar words, justifying that LSA bears resemblance to the human memory, more specifically to memory search and free recall (Zaromb et al. 2006; Landauer et al. 2007). Also, the evolution modeled through increasing corpora dimensions for deducing the word maturity metric (Landauer et al. 2011) underpins the cognitive similarities of word associations in terms of prior information or knowledge.

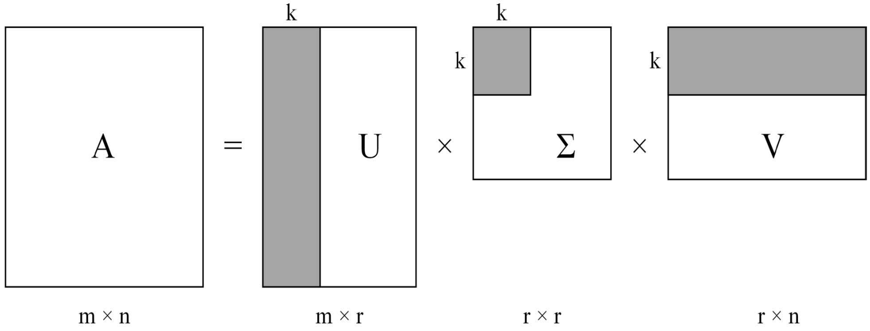


Fig. 10 Latent Semantic Analysis Decomposition (after Berry et al. 1995, p. 5)

From a computational perspective, LSA is used for evaluating the proximity between concepts or textual elements by cosine similarity or, equivalent, scalar product (see Equation 1). In addition to the initial model, multiple optimizations can be envisioned in order to increase the reliability of the semantic vector-space. Firstly, two crucial aspects, although empirical, need to be addressed: the initial document dimension and the number of dimension *k* after projection. In terms of documents size, semantically and topically coherent passages of approximately 50 to 100 words are the optimal units to be taken into consideration while building the initial matrix (Landauer and Dumais 2008). While considering the number of dimensions *k*, 300 can be considered an optimal empiric value agreed by multiple sources (Berry et al. 1999; Lemaire 2009; Landauer et al. 2007; Jessup and Martin 2001; Lizza and Sartoretto 2001).

$$Sim(word_1, word_2) = \frac{\sum_{i=1}^{k} word_{1,i} * word_{2,i}}{\sqrt{\sum_{i=1}^{k} word_{1,i}^2} * \sqrt{\sum_{i=1}^{k} word_{2,i}^2}} \quad (1)$$

Secondly, term weighting (Dumais 1991) can be applied on the elements of the initial term-document matrix. Term frequency – inverse document frequency (*Tf-Idf*) (Manning and Schütze 1999) provides a practical approach due to its duality: 1/ local importance, reflected in the normalization of the number of appearances of a word in a given document and 2/ global significance by weighting the appearances of a given word in all corpus documents, therefore enhancing the importance of rare words and reducing the significance of common ones (see Equation 2). Moreover, although word vectors can be directly summed up in order to build the representation of larger textual fragments, normalization of contained concepts also improves overall performance.

$$w_{D,i} = \left(\ln\left(tf_{D,i} + 1\right)\right) * ln\frac{N}{n_i} \quad (2)$$

where $tf_{D,i}$ is the number of occurrences of the term *i* in document *D*, *N* is the total number of documents in the corpus and nn_{i} is the number of documents in which the term *i* occurs.

Thirdly, POS tagging (Wiemer-Hastings and Zipitria 2001; Rishel et al. 2006) can be applied on all remaining words after stop word elimination and all inflected forms can be reduced to their lemma (Dascalu et al. 2010c; Bestgen 2012), that means enforcing the NLP pipe on the training corpus. According to Lemaire (2009) and Wiemer-Hastings and Zipitria (2001), stemming applied on all words reduces overall performance because each inflected form can expresses different perceptions and is related to different concepts. Therefore, as compromise of all previous NLP specific treatments, the latest version of the implemented tagged LSA model (Dascalu et al. 2013a; Dascalu et al. 2013b) uses lemmas plus their corresponding part-of-speech, after initial input cleaning and stop words elimination. In the end, due to the high demand of computational resources when performing the SVD decomposition on a sparse matrix of at least 20K terms with 20K passages (Landauer and Dumais 2008) (see *7.2* Cohesion-based Discourse Analysis), distributed computing enabling a concurrent and parallel execution of tasks can be considered a necessity for increasing speedup.

Similar to semantic distances, we must also consider the limitations of LSA, correlated to the experiments performed by Gamallo and Bordag (2011): 1/ the requirement of a large corpus of documents for training, both domain specific and general; 2/ the computational constraints due to the SVD decomposition phase; 3/ the model is blind to word order and to polysemy, as all word senses are merged into a single concept; 4/ the empirical selection of *k* and the segmentation of the initial documents into cohesive units of a given size, although cooccurrence patterns emerge in large training corpora; and 5/ despite the fact that updating mechanisms have been devised for increasing the training corpora (Berry et al. 1995; Witter and Berry 1998), it is unfeasible to apply them in practice, and once trained, the model remains unchanged.

4.3.3 Topic Relatedness through Latent Dirichlet Allocation

The goal of Latent Dirichlet Allocation (LDA) topic models is to provide an inference mechanism of underlying topic structures through a generative probabilistic process (Blei et al. 2003). Starting from the presumption that documents integrate multiple topics, each document can now be considered a random mixture of corpus-wide topics. In order to avoid confusion, an important aspect needs to be addressed: topics within LDA are latent classes, in which every word has a given probability, whereas topics that are identified within subsequently developed systems (*A.S.A.P.*, *Ch.A.M.P.*, *PolyCAFe* and *ReaderBench*) are key concepts from the text. Additionally, similar to LSA, LDA also uses the implicit assumption of the bag of words approach that the order of words doesn't matter when extracting key concepts and similarities of concepts through co-occurrences within a large corpus. In contrast to LSA (Landauer 2002) and *WordNet* (Miller 1998) that have empirically proved cognitive bases, LDA does not have such a cognitive argumentation; it is a probabilistic topic model in which the connotations of the latent space behind the model can be ignored (Chang et al. 2009).

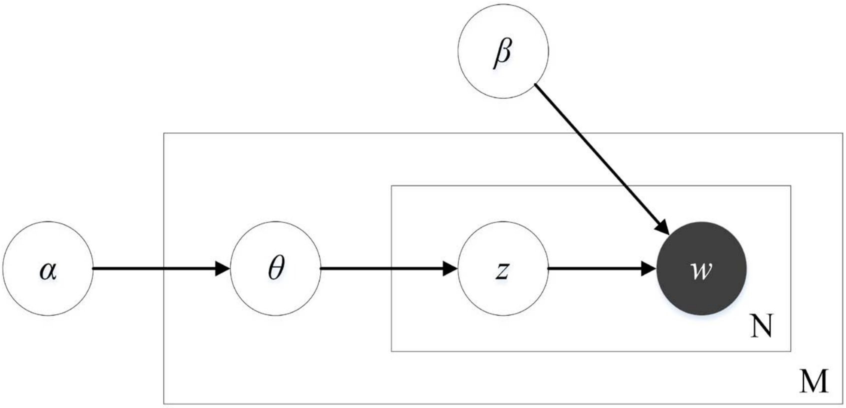


Fig. 11 Latent Dirichlet Allocation – graphical model representation (after Blei et al. 2003, p. 997). $w_{d,n}$ – n^{th} observed word in d document; $z_{d,n}$ – per word topic assignment; θ_d – per document topic proportions; β_k – per corpus topics distributions; M – corpus of documents; α – Dirichlet parameter; Each structure can be considered a random variable.

Every topic contains a probability for every word, but after the inference phase a remarkable demarcation can be observed between salient or dominant concepts of a topic and all other vocabulary words. In other words, the goal of LDA is to reflect the thematic structure of a document or of a collection through hidden variables and to infer this hidden structure by using a posterior inference model (Blei et al. 2003) (see Figure 11). Later on, as documents can be considered a mixture of topics, LDA focuses on situating new documents in the estimated pre-trained model. A topic is a Dirichlet distribution (Kotz et al. 2000) over the vocabulary simplex (the space of all possible distributions of words from the training corpora) in which thematically related terms have similar probabilities of occurrences. Moreover, as the Dirichlet parameter can be used to control sparsity, penalizing a document for using multiple topics, LDA's topics reflect in the end sets of concepts that co-occur more frequently (Blei and Lafferty 2009).

Therefore, documents become topics distributions drawn from Dirichlet distributions and similarities between textual fragments can be expressed by comparing the posterior topic distributions. Due to the fact that KL divergence (see Equation 3) (Kullback and Leibler 1951) is not a proper distance measure, as it is not symmetric, Jensen-Shannon dissimilarity (see Equation 4) (Manning and Schütze 1999; Cha 2007) can be used as a smoothed, symmetrized alternative. In the end, semantic similarity between textual fragments can be computed in terms of relatedness between distributions of topics – $prob(fragment_i)$, more specifically the inverse of the Jensen-Shannon distance (see Equation 5):

$$D_{KL}(P||Q) = \sum_i \left(\frac{P(i)}{Q(i)}\right) P(i) \tag{3}$$

$$D_{JS}(P||Q) = \frac{1}{2}(D_{KL}(P||M) + D_{KL}(Q||M)), M = \frac{1}{2}(P + M) \tag{4}$$

$$sim(fragment_1, fragment_2) = 1 - D_{JS}(prob(fragment_1), prob(fragment_2)) \tag{5}$$

Despite the fact that LDA uses only few latent variables, exact inference is generally intractable (Heinrich 2008). Therefore, the solution consists of using approximate inference algorithms, from which Gibbs sampling (Griffiths 2002) seems most appropriate and is most frequently used. Gibbs sampling can be considered a special case of Markov-chain Monte Carlo (MCMC) simulation (MacKay 2003) and integrates relatively simple algorithms for approximating inference in high-dimensional models (Heinrich 2008) – k, the number of topics, is usually 100, as suggested by Blei et al. (2003). Of particular interest from a computational point of view is the possibility to perform a distributed Gibbs sampling (McCallum 2002) in order to increase training speedup.

Although LDA proved to be reliable in extracting topics and has the lowest perplexity levels (a measure algebraically equivalent to the inverse of the geometric mean per-word likelihood) when compared to other probabilistic semantic models (Blei et al. 2003), we must also consider its drawbacks: 1/ although topics reflect terms that more tightly co-occur, there are no actual class significances automatically deduced and topics are not equi-probable (Arora and Ravindran 2008); 2/ by using an approximate inference model, there are inevitably estimation errors, more notable when addressing smaller documents or texts with a wider spread of concepts, as the mixture of topics becomes more uncertain; 3/ similarly to LSA, LDA is blind to word order, but polysemy is reflected in the membership of the same word, with high probabilities, in multiple topics; and 4/ LDA, in comparison to LSA, loses the cognitive significance and the posterior distributions are nevertheless harder to interpret than the semantic vector space representations of concepts.

References

Adams, P.H., Martell, C.H.: Topic Detection and Extraction in Chat. In: IEEE Int. Conf. on Semantic Computing (ICSC 2008), pp. 581–588. IEEE, Santa Clara (2008)

Agirre, E., Lopez, O.: Clustering WordNet Word Senses. In: Conference on Recent Advances on Natural Language (RANLP 2003), pp. 121–130. ACL, Borovetz (2003)

Arora, R., Ravindran, B.: Latent dirichlet allocation based multi-document summarization. In: 2nd Workshop on Analytics for Noisy Unstructured Text Data, Singapore, pp. 91–97. ACM (2008), doi:10.1145/1390749.1390764

Bakhtin, M.M.: The dialogic imagination: Four essays (trans: Emerson C, Holquist M). The University of Texas Press, Austin (1981)

Bakhtin, M.M.: Problems of Dostoevsky's poetics (trans: Emerson C). University of Minnesota Press, Minneapolis (1984)

Bakhtin, M.M.: Speech genres and other late essays (trans: McGee VW). University of Texas, Austin (1986)

Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 136–145. Springer, Heidelberg (2002)

Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: ACL Workshop on Intelligent Scalable Text Summarization (ISTS 1997), pp. 10–17. ACL, Madrid (1997)

Berry, M.W., Drmac, Z., Jessup, E.R.: Matrices, vector spaces, and information retrieval. SIAM Review 41(2), 335–362 (1999)

Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. SIAM Review 37, 573–595 (1995)

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

Blei, D.M., Lafferty, J.: Topic Models. In: Srivastava, A., Sahami, M. (eds.) Text Mining: Classification, Clustering, and Applications. CRC Data Mining and Knowledge Discovery, pp. 71–93. Chapman & Hall/CRC, London (2009)

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

Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA, pp. 29–34 (2001)

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

Cha, S.H.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007)

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) 23rd Annual Conference on Neural Information Processing Systems (NISP 2009), Vancouver, Canada, 2009, pp. 288–296 (2009)

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. (eds.): Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)

Cree, G.S., Armstrong, B.C.: Computational Models of Semantic Memory. In: Spivey, M., McRae, K., Joanisse, M. (eds.) The Cambridge Handbook of Psycholinguistics, pp. 259–282. Cambridge University Press, Cambridge (2012), http://dx.doi.org/10.1017/CBO9781139029377.018

Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., Nardy, A.: ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 379–388. Springer, Heidelberg (2013)

Dascalu, M., Trausan-Matu, S., Dessus, P.: Utterances assessment in chat conversations. Research in Computing Science 46, 323–334 (2010c)

Dascalu, M., Trausan-Matu, S., Dessus, P.: Cohesion-based Analysis of CSCL Conversations: Holistic and Individual Perspectives. In: Rummel, N., Kapur, M., Nathan, M., Puntambekar, S. (eds.) 10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013), pp. 145–152. ISLS, Madison (2013b)

Deerwester, S., Dumais, S.T., Furnas, G.W., Harshman, R., Landauer, T.K., Lochbaum, K., Streeter, L.: Computer information retrieval using latent semantic structure USA Patent 4,839,853 (June 13, 1989)

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)

Dong, A.: The latent semantic approach to studying design team communication. Design Studies 26(5), 445–461 (2005)

Dong, A.: Concept formation as knowledge accumulation: A computational linguistics study. AIE EDAM: Artificial Intelligence for Engineering Design, Analysis, and Manufacturing 20(1), 35–53 (2006)

Dumais, S.T.: Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers 23(2), 229–236 (1991)

Dumais, S.T.: Latent semantic analysis. Annual Review of Information Science and Technology 38(1), 188–230 (2004)

Emond, B.: WN-LEXICAL: An ACT-R module built from the WordNet lexical database. In: 11th Int. Conf. on Cognitive Modeling, Trieste, Italy, 2006, pp. 359–360 (2006)

Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)

Fellbaum, C.: WordNet(s). In: Brown, K. (ed.) Encyclopedia of Language and Linguistics, 2nd edn., vol. 13, pp. 665–670. Elsevier, Oxford (2005)

Foltz, P.W., Kintsch, W., Landauer, T.K.: An analysis of textual coherence using latent semantic indexing. In: 3rd Annual Conference of the Society for Text and Discourse, Boulder, CO (1993)

Foltz, P.W., Kintsch, W., Landauer, T.K.: The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes 25(2-3), 285–307 (1998)

Galley, M., McKeown, K.: Improving Word Sense Disambiguation in Lexical Chaining. In: Gottlob, G., Walsh, T. (eds.) 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), pp. 1486–1488. Morgan Kaufmann Publishers, Inc., Acapulco (2003)

Gamallo, P., Bordag, S.: Is singular value decomposition useful for word similarity extraction? Language Resources and Evaluation 45(2), 95–119 (2011)

Genesereth, M.R., Nilsson, N.J.: Logical Foundations of Artificial Intelligence. Morgan Kaufmann Publishers, San Mateo (1987)

Golub, G.H., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis 2(2), 205–224 (1965)

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

Gordon, P.C., Grosz, B.J., Gillom, L.A.: Pronouns, Names and the Centering of Attention in discourse. Cognitive Science 17(3), 311–347 (1993)

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavioral Research Methods, Instruments, and Computers 36(2), 193–202 (2004)

Griffiths, T.: Gibbs sampling in the generative model of Latent Dirichlet Allocation. Stanford University, Stanford (2002)

Grosz, B.J., Sidner, C.L.: Attention, intentions, and the structure of discourse. Computational Linguistics 12(3), 175–204 (1986)

Grosz, B.J., Weinstein, S., Joshi, A.K.: Centering: a framework for modeling the local coherence of discourse. Computational Linguistics 21(2), 203–225 (1995)

Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)

Halliday, M.A.K., Hasan, R.: Cohesion In English. Longman, London (1976)

Heinrich, G.: Parameter estimation for text analysis. vsonix GmbH + University of Leipzig, Leipzig (2008)

Hirst, G., St-Onge, D.: Lexical Chains as representation of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, p. 25. MIT Press, Cambridge (1997)

Hobbs, J.R.: Why is Discourse Coherent? SRI International, Menlo Park (1978)

Hobbs, J.R.: Coherence and Coreference. Cognitive Science 3(1), 67–90 (1979)

Hobbs, J.R.: On the Coherence and Structure of Discourse. Stanford University, Center for the Study of Language and Information (1985)

Hobbs, J.R.: Topic drift. In: Dorval, B. (ed.) Conversational Organization and its Development, pp. 3–22. Ablex Publishing Corp., Norwood (1990)

Holmer, T., Kienle, A., Wessner, M.: Explicit Referencing in Learning Chats: Needs and Acceptance. In: Nejdl, W., Tochtermann, K. (eds.) EC-TEL 2006. LNCS, vol. 4227, pp. 170–184. Springer, Heidelberg (2006)

Howard, M.W., Kahana, M.J.: Temporal Associations and Prior-List Intrusions in Free Recall. Journal of Experimental Psychology: Learning, Memory, and Cognition 25(4), 923–941 (1999)

Jessup, E.R., Martin, J.H.: Taking a new look at the Latent Semantic Analysis approach to information retrieval. In: Berry, M.W. (ed.) Computational Information Retrieval, pp. 121–144. SIAM, Philadelphia (2001)

Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Int. Conf. on Research in Computational Linguistics (ROCLING X), pp. 19–33. Academia Sinica, Taipei (1997)

Joshi, M., Rosé, C.P.: Using Transactivity in Conversation Summarization in Educational Dialog. In: SLaTE Workshop on Speech and Language Technology in Education, Farmington, Pennsylvania, USA (2007)

Jurafsky, D., Martin, J.H.: An introduction to natural language processing. Computational linguistics, and speech recognition, 2nd edn. Pearson Prentice Hall, London (2009)

Kintsch, W.: Metaphor comprehension: A computational theory. Psychonomic Bulletin and Review 7, 257–266 (2000)

Kintsch, W.: Predication. Cognitive Science 25(2), 173–202 (2001)

Kontostathis, A., Edwards, L., Bayzick, J., McGhee, I., Leatherman, A., Moore, K.: Comparison of Rule-based to Human Analysis of Chat Logs. In: Meseguer, P., Mandow, L., Gasca, R.M. (eds.) 1st International Workshop on Mining Social Media Programme, Conferencia de la Asociación Española Para La Inteligencia Artificial, Seville, Spain, p. 12. Springer (2009)

Koschmann, T.: Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration. In: Hoadley, C.M., Roschelle, J. (eds.) Int. Conf. on Computer Support for Collaborative Learning (CSCL 1999), pp. 308–313. ISLS, Palo Alto (1999)

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

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

Landauer, T.K.: On the computational basis of learning and cognition: Arguments from LSA. The Psychology of Learning and Motivation 41, 43–84 (2002)

Landauer, T.K., Dumais, S.: Latent semantic analysis. Scholarpedia 3(11), 4356 (2008)

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

Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to Latent Semantic Analysis. Discourse Processes 25(2/3), 259–284 (1998a)

Landauer, T.K., Kireyev, K., Panaccione, C.: Word maturity: A new metric for word knowledge. Scientific Studies of Reading 15(1), 92–108 (2011)

Landauer, T.K., Laham, D., Foltz, P.W.: Learning human-like knowledge by singular value decomposition: a progress report. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems, vol. 10, pp. 45–51. MIT Press, Cambridge (1998b)

Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.): Handbook of Latent Semantic Analysis. Erlbaum, Mahwah (2007)

Lapata, M., Barzilay, R.: Automatic evaluation of text coherence: models and representations. In: 19th International Joint Conference on Artificial Intelligence, pp. 1085–1090. Morgan Kaufmann Publishers Inc., Edinburgh (2005)

Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for wordsense identification. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)

Lemaire, B.: Limites de la lemmatisation pour l'extraction de significations. In: 9es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2009), pp. 725–732. Presses Universitaires de Lyon, Lyon (2009)

Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: DeBuys, V. (ed.) 5th Annual Int. Conf. on Systems documentation (SIGDOC 1986), pp. 24–26. ACM, Toronto (1986)

Lin, D.: An information-theoretic definition of similarity. In: 15th Int. Conf. on Machine Learning, pp. 296–304. Morgan Kaufmann, Madison (1998)

Lizza, M., Sartoretto, F.: A comparative analysis of LSI strategies. In: Berry, M.W. (ed.) Computational Information Retrieval, pp. 171–181. SIAM, Philadelphia (2001)

MacKay, D.J.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)

Mahnkopf, C.S.: Theory of Polyphony. In: Mahnkopf, C.S., Cox, F., Schurig, W. (eds.) Polyphony and Complexity, vol. 1, p. 328. Wolke Verlags Gmbh, Hofheim (2002)

Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: A Theory of Text Organization. Information Sciences Institute, Marina del Rey (1987)

Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8(3), 243–281 (1988)

Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)

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

McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit. University of Massachusetts Amherst, Amherst (2002)

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C.: Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes 47(4), 292–330 (2010)

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

Miller, G.A.: Foreword. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. xv–xxii. MIT Press, Cambridge (1998)

Miller, G.A.: WordNet. Princeton University Press, Princeton (2010)

Miltsakaki, E., Kukich, K.: The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In: 38th Annual Meeting on Association for Computational Linguistics, pp. 408–415. ACL, Hong Kong (2000)

Moldovan, D.I., Mihalcea, R.: Using WordNet and Lexical Operators to Improve Internet Searches. IEEE Internet Computing 4(1), 34–43 (2000)

Moore, J.D., Pollack, M.E.: A problem for RST: the need for multi-level discourse analysis. Computational Linguistics 18(4), 537–544 (1992)

Morris, J., Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. Computational Linguistics 17(1), 21–48 (1991)

Murphy, M.L.: Semantic Relations and the Lexicon: antonymy, synonymy and other paradigms. Cambridge University Press, Cambridge (2003)

Navigli, R.: Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In: 44th Annual Meeting of the Association for Computational Linguistics Joint with the 21st Int. Conf. on Computational Linguistics (COLING-ACL 2006), pp. 105–112. ACL, Sydney (2006)

Navigli, R.: Word Sense Disambiguation: A Survey. ACM Computing Surveys 41(2), 1–69 (2009)

Navigli, R., Velardi, P.: An Analysis of Ontology-based Query Expansion Strategies. In: Workshop on Adaptive Text Extraction and Mining (ATEM 2003), in the 14th European Conference on Machine Learning (ECML 2003), Cavtat-Dubrovnik, Croatia, 2003, pp. 42–49. Springer (2003)

Oltramari, A., Gangemi, A., Guarino, N., Masolo, C.: Restructuring WordNet's Top-Level: The OntoClean approach. In: OntoLex'2 Workshop, Ontologies and Lexical Knowledge Bases (LREC 2002), Las Palmas, Spain, 2002, pp. 17–26 (2002)

Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet:Similarity -Measuring the Relatedness of Concepts. In: Nineteenth National Conference on Artificial Intelligence (AAAI 2004), San Jose, CA, pp. 1024–1025 (2004)

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.A.: Multi-Pass Sieve for Coreference Resolution. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), pp. 492–501. ACL, Cambridge (2010)

Resnik, P.: Using information content to evaluate semantic similarity. In: 14th International Joint Conference on Artificial Intelligence, pp. 448–453. Morgan Kaufmann, Montreal (1995)

Rishel, T., Perkins, A.L., Yenduri, S., Zand, F.: Augmentation of a Term/Document Matrix with Part-ofSpeech Tags to Improve Accuracy of Latent Semantic Analysis. In: 5th WSEAS Int. Conf. on Applied Computer Science, Hangzhou, China, pp. 573–578 (2006)

Rosé, C.P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning. International Journal of Computer Supported Collaborative Learning 3(3), 237–271 (2008)

Sagot, B.: WordNet Libre du Francais (WOLF). INRIA, Paris (2008)

Sagot, B., Darja, F.: Building a free French wordnet from multilingual resources. In: Ontolex 2008, Marrakech, Maroc, p. 6 (2008)

Schmidt, A.P., Stone, T.K.M.: Detection of Topic Change in IRC Chat Logs (2013), `http://www.trevorstone.org/chatsegmentation/`

Silber, G., McCoy, K.: Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics -Summarization 28(4), 487–496 (2003)

Snow, R., Prakash, S., Jurafsky, D., Ng, A.Y.: Learning to Merge Word Senses. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp. 1005–1014. ACL, Prague (2007)

Stahl, G.: Studying Virtual Math Teams. Springer, New York (2009a)

Trausan-Matu, S.: The Polyphonic Model of Hybrid and Collaborative Learning. In: Wang, F., Fong, L. (eds.) J, Kwan RC (eds) Handbook of Research on Hybrid Learning Models: Advanced Tools, Technologies, and Applications, pp. 466–486. Information Science Publishing, Hershey (2010c)

Trausan-Matu, S.: Experiencing, Conducting, Designing and Evaluating Polyphony in CSCL Chats. In: Spada, H., Stahl, G., Miyake, N., Law, N. (eds.) 9th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2011), pp. 906–907. ISLS, Hong Kong (2011)

Trausan-Matu, S.: From Two-Part Inventions for Three Voices, to Fugues and Creative Discourse Building in CSCL Chats. Unpublished manuscript (2013)

Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 352–357. Springer, Heidelberg (2012)

Trausan-Matu, S., Rebedea, T.: Polyphonic Inter-Animation of Voices in VMT. In: Stahl, G. (ed.) Studying Virtual Math Teams, pp. 451–473. Springer, Boston (2009)

Trausan-Matu, S., Rebedea, T.: A Polyphonic Model and System for Inter-animation Analysis in Chat Conversations with Multiple Participants. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 354–363. Springer, Heidelberg (2010)

Trausan-Matu, S., Rebedea, T., Dascalu, M.: Analysis of discourse in collaborative Learning Chat Conversations with Multiple Participants. In: Tufis, D., Forascu, C. (eds.) Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, pp. 313–330. Editura Academiei, Bucharest (2010b)

Trausan-Matu, S., Rebedea, T., Dragan, A., Alexandru, C.: Visualisation of learners' contributions in chat conversations. In: Fong, J., Wang, F.L. (eds.) Blended Learning, pp. 217–226. Pearson/Prentice Hall, Singapour (2007a)

Trausan-Matu, S., Stahl, G.: Polyphonic inter-animation of voices in chats. In: CSCL2007 Workshop on Chat Analysis in Virtual Math Teams, p. 12. ISLS, New Brunwick (2007)

Trausan-Matu, S., Stahl, G., Sarmiento, J.: Polyphonic Support for Collaborative Learning. In: Dimitriadis, Y.A., Zigurs, I., Gómez-Sánchez, E. (eds.) CRIWG 2006. LNCS, vol. 4154, pp. 132–139. Springer, Heidelberg (2006)

Trausan-Matu, S., Stahl, G., Sarmiento, J.: Supporting polyphonic collaborative learning. E-service Journal 6(1), 58–74 (2007b)

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

Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Croft, W.B. (ed.) 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994), Dublin, Ireland, pp. 61–69. Springer (1994)

Wang, T., Hirst, G.: Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), pp. 1003–1011. ACL, Edinburgh (2011)

Wegerif, R.A.: dialogical understanding of the relationship between CSCL and teaching thinking skills. In: Koschmann, T., Suthers, D., Chan, T.W. (eds.) Conf. on Computer Supported Collaborative Learning 2005: The Next 10 Years (CSCL 2005), p. 7. ISLS, Taipei (2005)

Wiemer-Hastings, P., Zipitria, I.: Rules for syntax, vectors for semantics. In: 23rd Annual Conference of the Cognitive Science Society, pp. 1112–1117. Erlbaum, Mahwah (2001)

Witter, D., Berry, M.W.: Downdating the latent semantic indexing model for conceptual information retrieval. The Computer Journal 41, 589–601 (1998)

Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, ACL 1994, pp. 133–138. ACL, New Mexico (1994)

Yeh, J.-F., Wu, C.-H., Chen, M.-J.: Ontology-based speech act identification in a bilingual dialog system using partial pattern trees. Journal of the American Society for Information Science and Technology 59(5), 684–694 (2008)

Zaromb, F.M., Howard, M.W., Dolan, E.D., Sirotin, Y.B., Tully, M., Wingfield, A., Kahana, M.J.: Temporal Associations and Prior-List Intrusions in Free Recall. Journal of Experimental Psychology: Learning, Memory, and Cognition 32(4), 792–804 (2006)