Real-time Analysis of Speech Streams and their Representation as Conceptual Structures

Ronny Melz, Chris Biemann, Karsten Böhm, Gerhard Heyer, Fabian Schmidt

University of Leipzig, Institute of Computer Science
Augustusplatz 10-11, 04109 Leipzig, Germany
[rmelz, biem, boehm, heyer, fschmidt]@informatik.uni-leipzig.de

Abstract

Explicit semantic context directly furthers understanding of content in humans. Applications range from efficient information exchange in meetings to visual information retrieval or orientation in document collections. The automatically extended context addresses various issues simultaneously: a problem is depicted from different points of view, new ideas are associated to those already mentioned, the trail lends an assisting and moderating functionality to a discussion and finally the network of concepts can be saved or printed out for later reference. Obviously, discussions only benefit from this context, if provided in real-time. In this contribution we present the software "*SemanticTalk*", which provides access to the underlying conceptual structures of unstructured speech streams. Essentially, the most significant concepts of a speech stream are extracted and visualized and related to further associated concepts automatically retrieved from a collection of documents. This results in an incrementally growing conceptual graph, which is dynamically unfolded in real-time.

1 Introduction

Depending on professional background and common environment, each human being has its own definitions and correlations of concepts, a so-called "semantic space". On the one hand, a concept may significate different things and hence be related to different concepts, on the other hand, various words may correspond to the same concept. Only within specific, rather homogeneous interest groups – such as either designers, managers or engineers – concepts usually bear the same meaning.

Meetings like consortia, planning assemblies, strategic discussions or brainstorming sessions still belong to the most important means of cooperative interaction. These highly collaborative environments require interdisciplinary knowledge and discourse in particular. In project meetings, the expert knowledge of each domain is of essential importance for the success of the proceeding. Hence the heterogeneous semantic spaces hinder efficient communication: the focus of a discussion could get lost in details, members unconsciously suppose important clues to be known and forget them to explain, or certain members could get confused by polysemous concepts.

SemanticTalk (www.isa.de/produkte/SemanticTalk) is a software which has been developed to display conceptual structures. Two instances of its main application are live-assistance of strategic meetings in business and visalization of conceptual maps to serve as an overview over a certain domain. Optionally, the concepts from another input stream can be embedded into the overview map to reveal its most relevant concepts along a "red thread" – this facilitates recognition of contents coverage. The applications are described in (Biemann, Böhm, Heyer & Melz, 2004a) and (Biemann, Heyer, Schmidt & Witschel, 2004b).

Since the prototype state of the above mentioned citations, the system was continually developed into a sophisticated conceptual platform. Apart the multi-user inputs of free speech, of digital text and precalculated semantic maps, the software now provides a powerful bidirectional XML interface to be integrated into virtually all standard software environments such as the Aris process platform (http://www.ids-scheer.com/international/english/products/31207). Furthermore, functionality has been extended to actively steer and manipulate the visualized graphs.

This paper focuses on the technical issues and implementation of the software system. The remaining content is organized as follows: Section 2 describes the architecture and the underlying techniques in detail. Section 3 outlines the graphical user interface and its broad functionality. Section 4 sketches the extended application domain and section 5 finally terminates with some concluding remarks and further research directions.

2 Architecture of the System and Methodology

Although the system can easily be set up on a single PC, it is implemented as a platform-independent client-server architecture. Several GUIs as well as independent speaker input units, consisting of a standard microphone and a client notebook, can be locally distributed and connected via standard networks like WLAN. Hence, the motive independence during meetings is largely conserved.

2.1 Interfaces

SemanticTalk provides various interfaces to both data sources and sinks, the most important of which are doubtlessly the input module for free speech, the graph distiller to automatically assemble static semantic maps and the XML-based interface to bidirectionally exchange structured data.

2.1.1 Speech Input

For speech-to-text conversion, we use a commercially available dictation system, Linguatec's VoicePro©. Since VoicePro© has not been designed originally for the use in multi-user environments, the decoder cannot reach its maximum accurracy. Nevertheless, the performance is noteworthy and the generated word stream is more than sufficient for our purposes. As mentioned above, *SemanticTalk* is operable as a distributed system. A further advantage of this approach is, that the comparatively high computational burden of natural speech decoding can be shifted to the clients.

2.1.2 Graph Distiller

The graph distiller takes one or more text documents as input and generates a corresponding conceptual view on the data. The interconnection of concepts is based on the same principles as described in section 2.3. During distilling, no further concepts are associated, the output is a network of only the characteristic, domain specific concepts and their relationship. However, there is a significant difference to section 2.2. Section 2.2 describes, how key concepts are extracted from an incoming sequence of words. This sequence is characterized by the property to be causal: at some given time point t_0, only the present word and the history of words is available to be added to the graph. The future of the process is not determined. The graph distiller by contrast benefits from the availability of the whole text to construct the graph and further from the boundary condition, that it is not required to work in real-time. Hence, a more sophisticated algorithm is applied to extract the relevant concepts.

The extraction is based on "Difference Analysis" (Faulstich, Quasthoff, Schmidt, Wolff, 2002): Based on pure language statistics, the distribution of terms in the given text documents is compared to the distribution of the same terms in the general use of a specified language, say German. If a term occurs significantly more often in the examined input text documents than in general German, it is called a 'key concept of the domain'. The conceptual graph is constructed of only these key concepts. As based on pure statistics, Difference Analysis is language independent. However, it requires a large, representative collection of digital documents. Since *SemanticTalk* was developed primarily for the German market, our reference corpus were the 'Wortschatz' databases, available at www.wortschatz.uni-leipzig.de.

Beyond Difference Analysis, the graph distiller optionally provides algorithms as described in (Witschel, 2004). For morphologically rich languages like German, complex noun decomposition and lemmatization particularly improve the output of the graph distiller.

2.1.3 XML Data

XML is the abbreviation for Extensible Markup Language and provides a widely accepted standard framework to exchange structured contents and metadata among applications. We provide an interface for Resource Description Format (RDF, cf. www.w3.org/TR/rdf-syntax-grammar), which is a subset of the XML language standard especially for representing graphs, their topology and features. Hence, virtually all kinds of graphs created in different applications can be loaded for further use in *SemanticTalk* through this interface. Moreover it is used to output graphs, which consequently can be processed by further existing software.

2.1.4 Other Interfaces

Several other modules cover the various document formats for unstructured text to improve the accessibility of SemanticTalk, especially the Graph Distiller. The visualization panel can be printed out in standard graphics format. The figure may serve as mind-map-like graph (Buzan, 1998) for later reference and remembrance of the addressed issues. A bidirectional interface to a broadly used database server facilitates the storage and organisation of graphs. Lastly, an additionial text input field, embedded into the GUI, serves various tasks like retrieving and focussing a certain concept from the graph.

2.2 Relevant Concepts

A crucial point is the filtering of the incoming word sequence from the speech input module: If we added each of the words to the evolving graph, there would be obviously no profit in overview. In Red-Thread-Mode (cf. Biemann et al., 2004a), this filtering is easy: a semantic map already visualizes a clustered graph of concept nodes. These nodes are characteristic for the previously analyzed domain (cf. section 2.1.2). Only concepts included in the map are extracted from the incoming word sequence. When visualizing content in Free-Association-Mode, no such background information is available. Secondly, the extraction cannot be based on an exhaustive Difference Analysis since an efficient algorithm is required which works in real-time. Lastly, the speech stream is causal and hence its future course is not known at some given time instant t_0. Nevertheless, a surprisingly simple filter strategy has proven to provide noteworthy results and still works efficiently: The first-order Markov approximation (Shannon, 1948, p. 7) yields the unigram structure of the words of a language. Although more sophisticated estimations are possible (Wendemuth, 2004, pp. 30-31; Jurafsky & Martin, 2000, chapter 6), this approximation serves well as a maximum-likelihood-estimation of term occurrence. Such probability estimations of term occurrence are easily computable given a representative, digital corpus of language.

Consequently, according to (Zipf, 1965), the frequency distribution associated to the terms of a language is not arbitrary. Basic terms, frequently used across all language domains and usually composed of only a single morpheme, rather serve to syntactically concatenate essential concepts. They reveal high redundancy and a low load of semantic content. On the other hand, rarely occurring concepts are usually characterized by high specificity and higher semantic content. They are familiar to only some few special domains[1]. Hence, corpus frequency can roughly be associated to semantic content, whereas we focus on the terms of medial frequency – neither too specialized, nor too general. The bandpass filtering is then made dependent on two empirical parameters: an upper and a lower frequency. To implement the upper bound, we compiled a list of the highest frequent terms of the used corpus to match the word stream against it and filter out least-significant words. An implementation of the lower bound is not necessary, since sparsely occurring terms do not have significant associations to other terms and thus they are filtered out automatically in the next step, which is described in the following subsection. Furthermore, this simple heuristics can easily be adapted to include different terms in the list or to make use of linguistic rules.

[1] In practice, due to the imperfectness of typing or OCR, the spelling mistakes of the corpus constitute another source of very rare terms. However, their frequencies remain low, if the corpus consists of enough different contributing sources, which usually is the case.

2.3 Interconnection and Automatic Association of Concepts

Humans think networked. Such conceptual networks are associative in the sense that concepts or ideas are associated stronger to each other, the more they are related in content. There are two essential types of relation, termed after (Saussure, 1916): two linguistic signs relate syntagmatically, if they supplement each other in content, such as "coffee" and "drink"; they relate paradigmatically, if they bear in common important semantic features, such as "dog" and "cat" (cf. also Heyer, Quasthoff & Wittig, 2005). Such associative conceptual networks are known to exist since Aristoteles and have been intensely investigated by semantic priming experiments in psychology (Harley, 2001, p. 156) and in the cognitive sciences (Steyvers & Tenenbaum, 2005). Our aim is to automatically calculate such conceptual networks from the relevant concepts of the speech stream and associate further relevant information to this network.

Many approaches have been pointed out to automatically recover the inherent semantic structure of human language. The most prominent among the language independent approaches are Latent Semantic Analysis (Landauer & Dumais, 1997) and Likelihood Ratio Tests (Dunning, 1993). However, we choose still a different approach based on the log-likelihood measure originally presented by (Krenn, 2000). The underlying assumption is quite intuitive: the more often two concepts appear together in the same context, the more they correlate semantically. Assuming independence, Two words w_A and w_B are expected to occur together with probability $P(w_A \,\&\, w_B) = P(w_A) \cdot P(w_B)$. 'Occurring together' requires a restricted language unit wherein this event may take place or not. This could be a window of constant size or sentences. In our analyses we included windows of size two (strong neighborhood) and whole sentences. Assuming further, that words only occur once per language unit, we can use the maximum-likelihood estimations to calculate the joint probability of a co-occurrence: $p = P(w_A \,\&\, w_B) = n_A/N \cdot n_B/N$, where n_A is the number of times, token w_A occurs in the corpus, and N is the amount of language units in the corpus. Obviously, the last assumption deviates significantly for highly frequent words like 'the' or 'he'. But luckily, these are not the co-occurrences of interest and they are pruned by the filtering as described in the previous subsection. The pdf (probability density function) of the co-occurrences is assumed a Poisson distribution with expectation value as well as standard deviation $\lambda = pN$. For the actually observed co-occurrences, $k > \lambda$ holds always and thus the pdf $P(X = k)$ is monotonously falling:

$$\partial_k P(X = k) = \partial_k \left(\frac{1}{k!} \lambda^k e^{-\lambda} \right) \Big|_{k>\lambda} < 0$$

Hence the inverse, $P(X = k)^{-1}$, can be used to rank the co-occurrence significance. This equation is not efficiently computable. However, since we do not require the significance to have the properties of a probability value, we just can use the logarithm of the probability for ranking due to the monotony of logarithm function (the base merely accounts for an arbitrary constant and is not significant for significance ranking).

$$sig(w_A, w_B) = \log_e P(X = k)^{-1} = -\ln \frac{1}{k!} \lambda^k e^{-\lambda}$$

For large k, the equation can further be simplified by applying Stirling's formula $k! \sim \sqrt{2\pi k}\,(k/e)^k$ and neglecting all non-contributing (independent of k) and vanishing terms. This results for large k in the 'significance formula' as an instance of the log-likelihood measure:

$$sig(w_A, w_B) = k \cdot (\ln k - \ln \lambda - 1)$$

This formula quantifies the intuition, that a relation should be rated more significant if actually occurring frequently (high k), but rated less significant, if expected to do so (high λ). Generally, relations of syntagmatic as well as paradigmatic character are extracted. The set of the most significant relations of a concept defines its global context (cf. Heyer, Quasthoff, Wittig, 2005). Not significant relations are dropped from the set depending on some empirical threshold. By this simplistic approach, the semantic environment of a word can be extracted automatically, resulting in co-occurrence graphs like the examples given below. Moreover, these methods are language independent as has been shown by (Biemann et al., 2004). Beyond this basic strategy, there are various methods to further enrich content such as co-occurrences of higher order. The essential idea is to regard co-occurrence sets itself as language units and then to iterate the basic process. Yet a further improvement requires advanced clustering algorithms or genuine human insight and manual effort. The algorithm is not designed to group the concepts or the extracted relations into distinct classes, but likewise additional knowledge can be easily subjoined in a second processing

stage: knowledge is available from digital thesauri like Roget's Thesaurus (Roget, 1946) or ontological resources like WordNet (Miller, 1995) and can be extended by induction rules as well as human annotators.

Co-occurrence analysis is time-intensive and requires once more a large corpus (digital collection of documents). Thus, it cannot be afforded in real-time. The corpus is therefore analyzed in advance and the relations among concepts are stored together with their significance into a relational database. The analysis is comparable to a 'lexical acquisition process', wherein concepts and their interrelations are learned. Hence, the database constitutes the 'conceptual memory' of *SemanticTalk*. After the extraction of a relevant concept from the speech stream (see previous subsection), the database is queried for its contiguity to those concepts which already had been mentioned. The most significant of these relations (based on some empirical threshold) are visualized (cf. Figure 1) together with the concepts in the next step (see following subsection).

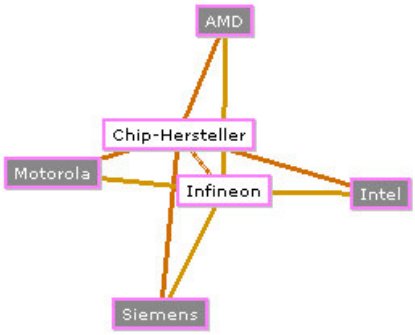

Figure 1: 'Chip-Hersteller' (chip manufacturer) and 'Infineon' are both extracted concepts. They are linked by a significant relation. Further semiconductor manufacturers are immediately associated (marked gray).

Beneath the automatic connection of conceptually related items from the speech stream, the database allows even for the association of new concepts, which had *not* been mentioned by the speakers! Since the database stores further relations than those necessary to link the actually spoken words, absent but still important concepts can easily be inferred from semantic context. For example, if from a speech stream "wo es eine **Uni** gibt sind auch **Studenten** nicht fern" ('usually, there are students around a university'), 'Uni' and 'Studenten' were extracted as relevant concepts, even this small context would allow for further human-like associations. In this example, *SemanticTalk* associates automatically 'Universität' and 'Professor'. The basic strategy is to check the 'global context set' of two or more concepts for common similar elements.

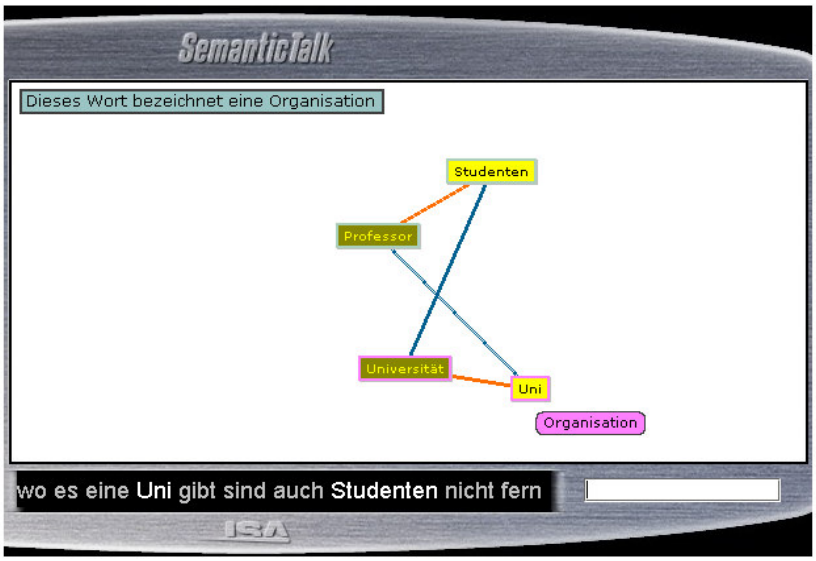

Figure 2: 'Uni' and 'Studenten' had been extracted as the most relevant concepts from the speech stream (see the text ticker at the bottom). Although there is no direct conceptual link, they are associated across 'Professor' and 'Universität', which *SemanticTalk* both added automatically though not mentioned explicitly.

2.4 Visualization

Formally, a network of interconnected concepts can be treated as a conceptual graph. A graph G is a mathematical object (more specifically, a pair $G(V_G, E_G)$) which consists of both a set of vertices and a set of edges. The (elements of the set of) vertices correspond to the concepts, hence the notation of a "conceptual graph". The (elements of the set of) edges correspond to semantic relations between these concepts. Our graphs are undirected, free of loops and do not contain multiple edges. Since the topology of a graph is invariant to any visual representation, there are various methods to actually draw graphs. Convenient for our purposes is the force-based model (Tollis, Di Battista, Eades & Tamassia, 1999, chapter 10). This model corresponds directly to the planar multi-body problem of theoretical mechanics: given a distribution of discrete bodies, their potential fields and the constraint forces acting amongst them, how does the position of the bodies evolve in time? The vertices are assigned an arbitrary (e.g. random) position on the plain. Next, each node is assigned a charge of the same sign. The modulus may be chosen to be a function of the node's properties (such as a weighting factor like the term frequency of the concept). However, we just assigned pairwise identical charges, just as if nodes were electrons. Lastly, each edge is assigned a spring constant k (which also could be based edge properties such as co-occurrence significance in our case). The algorithm now solves iteratively the change in resulting momentum by linear superposition of the attractive forces (due to the connecting springs) and repulsive forces (due to the potential field caused by the charges). Changes in position due to the momenta are tracked for each time step as shown in Figure 3. This way, the amount of energy in the system iteratively converges towards a local minimum (Koren, 2005).

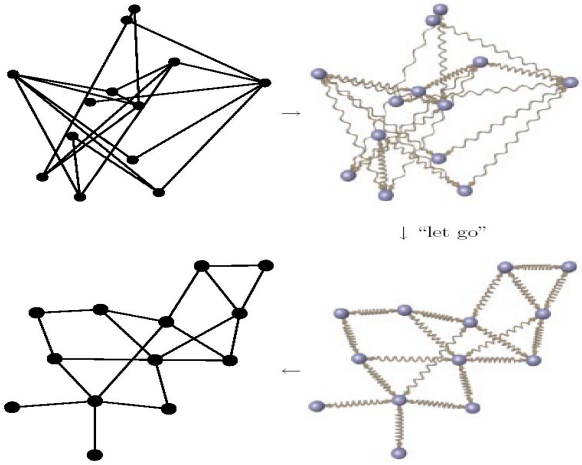

Figure 3: From random to order: by minimizing the global energy of the graph, any arbitrary initial configuration converges to an appealing visualization. In this model, nodes generate a repulsive potential field, whereas springs (edges) tie them together (Figure taken from Koren, 2005).

The force-directed model is convenient for our purposes as it provides an intuitive arrangement of the graph's concepts: For example, graph-inherent, partial symmetry is largely conserved (cf. Figure 1). More importantly, due to the global conservation of present structure for the most part as new concepts are associated, the incremental growth of the graph can be traced in its time-course. Here as well we need a time-efficient implementation to guarantee visualization in real-time. We use the publicly available "TouchGraph"-software (www.touchgraph.com). It can be obtained at prdownloads.sourceforge.net/touchgraph. The core classes of the TouchGraph package are enhanced by diverse functionality such as controls to steer the user's view on the data or additional XML-like attributes of nodes and edges. Due to various heuristic controls, virtually no human effort is required to steer the system while speaking freely.

3 Graphical User Interface

The main panel of *SemanticTalk* consists of the TouchGraph screen and shows the conceptual graph, which incrementally grows with each new extracted concept. The details have been described in the above section. Around the main panel, various other visualizations and controls are provided optionally, as shown in Figure 4.

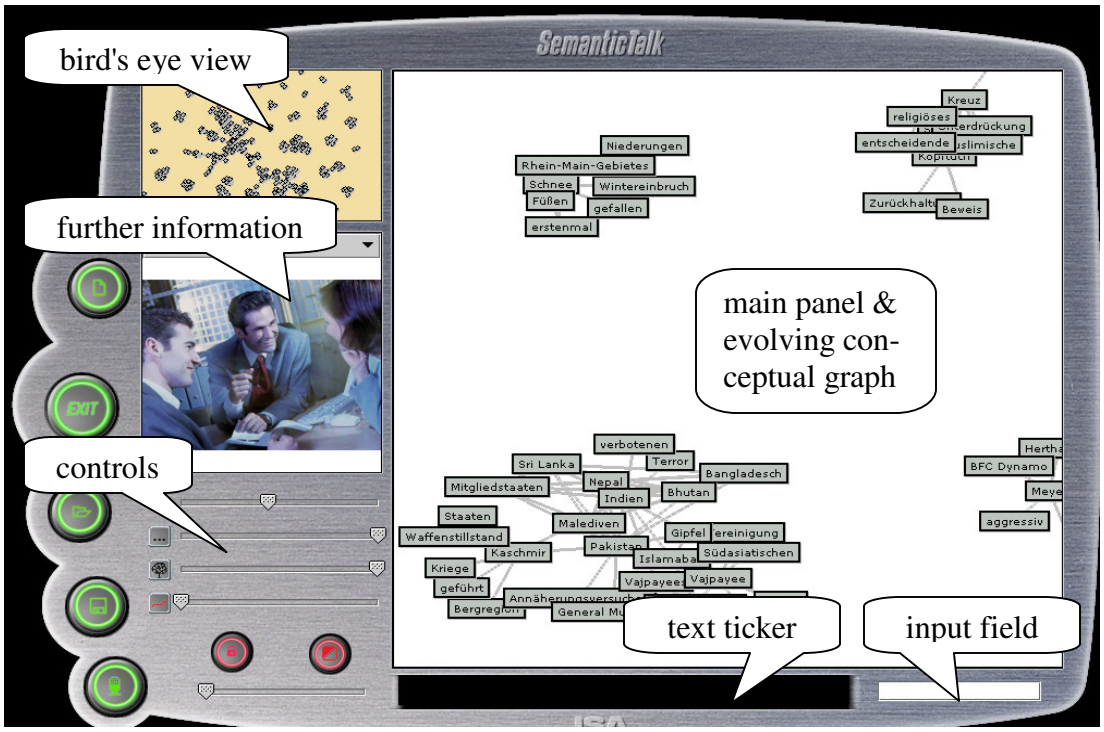

Figure 4: The GUI of *SemanticTalk* – the main panel contains the evolving conceptual graph. To the left there is an overview window, a field for further associated information and various controls. On the bottom, an input field is provided and the incoming word stream is echoed in the text ticker.

To the upper left, a smaller window appropriates a bird's eye view on the whole graph for overview, concept nodes are contracted into points. Directly thereunder, the currently extracted concepts from the speech stream are linked to further sources of information by direct access of the www or an annotated picture database, e.g. On the bottom of the GUI, a text ticker allows for the tracking of incoming speech: extracted concepts are emphasized by a lighter color; in Red Thread Mode, matching concepts are marked red. One of the functions of the generic input field is to focus some specific concept node of the graph just by typing its name. The leftmost buttons serve e.g. to load or save data such as graphs or text, or to switch the microphone on and off. The two remaining, smaller buttons switch between Free Association Mode and Red Thread Mode. Scrollbars control the zoom levels of the graph (cf. Fig. 5).

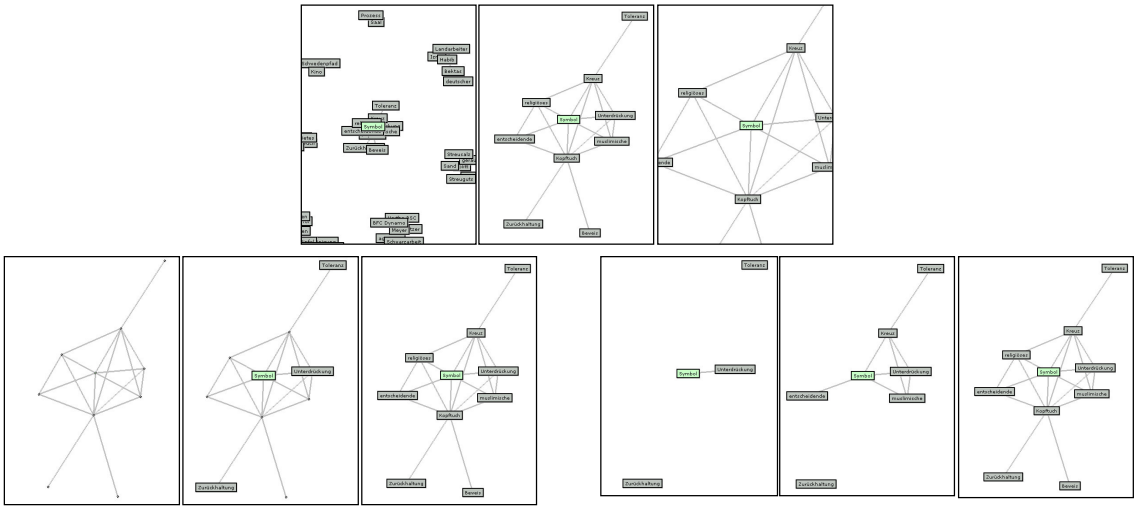

Figure 5: Different zoom methods – above the generic optical zoom, below on the left: the conceptual zoom, which either lexicalizes or contracts nodes to points based on their weight, to the right: the granularity zoom which hides or shows nodes. The latter zooms are based on node weight.

A last control which actively involves the user is the context menu. It pops up on right-clicking one of the graph's elements. Specific nodes for instance can be modified or further context can explicitly associated to them (cf. Fig.6).

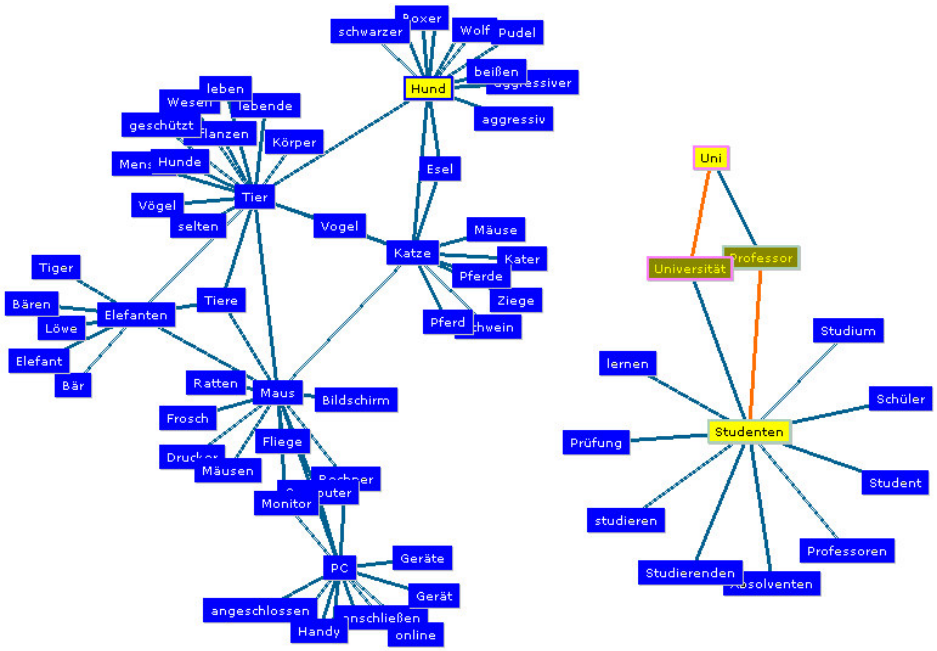

Figure 6: By right-clicking a concept, a context popup-menu allows for the explicit association of further concepts from the database virtually an arbitrary number of times. These concepts are marked blue here.

The aim of Free Association Mode is to assist an ongoing discussion and not to distract the participant's interest from the discussion's scope towards complex control-mechanisms. Hence efficient heuristics guess at convenient adjustments of the parameters such that practical no effort of any of the users is necessary. In Red Thread Mode, for instance, the focus follows the red thread and periodically zooms out for a few seconds to facilitate orientation also in larger graphs. Yet in Free Association Mode the locally fast changing regions of the graph are focused. Subsuming, *SemanticTalk* offers rich functionality to visualize, modify and exchange conceptual graphs, which are directly extracted from speech streams.

4 Example Applications

Our objective is to provide a platform for conceptual structures. Applications thereof are manifold. As has been pointed out by (Biemann et al., 2004a), there are essentially two modi: Free Association Mode and Red Thread Mode. In Free Association Mode, a conceptual graph is generated from scratch. Members of a discussion directly benefit from the evolving graph: it clearly reflects the covered themes of the discussion by concept clusters and thus assists to keep the discussion focused. Secondly, automatically associated context suggests ideas to be discussed more in detail, ambigue concepts, e.g. The third of the most important aspects of *SemanticTalk* is its 'creativity': particularly in brainstorming sessions, new ideas are stimulated when new concepts are associated to the graph. Assume an organization which wanted to extend its production to a new arising, specific market – such as an electronic display manufacturer who wanted to found a new division for large TFT panels. A brainstorming session together with managers, designers, engineers, is scheduled to get straight the concerned issues. Here, automatic associations of general language would be of less interest. But since the statistic methods of section 2.3 do not require anything more than large amounts of text, it is very easy to create associations, which are indeed very specific for the domain of interest: Large amounts of unstructured text are available from the internet, especially for new arising technologies. In the above mentioned example, this text would typically be downloaded by a simple script previous to the brainstorming session. By Difference and Cooccurrence Analysis, a domain specific database is compiled, which serves to extend the 'conceptual memory' of *SemanticTalk* during the actual session.

A second application is the generation and visualization of static conceptual graphs, which serve as orientation in document collections. Frequently, humans are required to extend their knowledge to hitherto unknown domains. With the always growing availability of electronic information, it is often difficult to grasp quickly just the most important concepts and their relations of a domain. *SemanticTalk* provides such an overview of a document collection: related key concepts are extracted and interlinked. An example is the map for the KnowTech conference (Biemann et al., 2004b). The generated semantic map describes efficiently the focus of each article and its relatedness to similarly focused contributions. While such semantic maps represent a whole domain, it is possible to display single documents as paths through them in Red Thread Mode: each concept extracted from input is retrieved in the pre-calculated semantic map, marked red, and linked to the last extracted concept by a red edge. The graph remains fixed (red edges do not apply further forces on the nodes). Hence the following can easily be examined:

coverage of domain: the more clusters are visited, the better the document covers the domain
coverage within topics: the more words in a cluster are marked in red, the more extensively the document deals with the topic of that cluster
relatedness of document and domain: few red nodes indicate non-relatedness
contiguity of document: many successive long range connections indicate semantic incoherence
visual comparison of different documents by using distinguishing colors

Since the path of a document immediately reveals its coverage of content with respect to some knowledge domain, this method speeds up a decision of whether to read a document or not to a few seconds.

5 Concluding Remarks and Future Work

SemanticTalk was first presented to public as a prototype implementation at CeBIT, March 2004. The final software release has been launched to market by ISA in December that year and since then a lots of positive user feedback encourages us on our way. At the moment, we are enriching the semantics of the XML-based connection to further support the embedding in standard business environments via Aris. The Aris Process Platform provides integrated tools for designing, implementing and controlling business processes, described by AML (Aris Modeling Language). The aim is to directly generate a business process model of the conceptual graph. Important additional functionality is the recognition of Aris objects and process steps. Vice versa, such models can be loaded and matched against a semantic map of special domain knowledge in Red Thread Mode, for instance. Finally, an adaption to various languages other than German is planned for the near future.

References

Biemann, C.; Böhm, K., Heyer, G. & Melz, R. (2004a). Automatically Building Concept Structures and Displaying Concept Trails for the Use in Brainstorming Sessions and Content Management Systems, Proceedings of I2CS, Guadalajara, Mexico and Springer LNCS

Biemann, C., Heyer, G., Schmidt, F. & Witschel, H.F. (2004b). Eine Wissenslandkarte der KnowTech, In N. Gronau, B. Petkoff, T. Schildhauer (Eds.), Wissensmanagement - Wandel, Wertschöpfung, Wachstum. (pp. 207-216). Berlin: GITO

Buzan, T. & Buzan, B. (1994). The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential, E P Dutton

Dunning, T.E. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp. 61-71

Faulstich, L., Quasthoff, U., Schmidt, F. & Wolff, C. (2002). Concept Extractor – Ein flexibler und domänen-spezifischer Web Service zur Beschlagwortung von Texten, ISI 2002

Harley, T. (2001). The Psychology of Language, Taylor & Francis Group

Heyer, G., Quasthoff, U. & Wittig, T. (2005). Wissensrohstoff Text, subject to appear at W3L-Verlag, Dortmund

Koren, Y. (2005). Drawing Large Graphs, www.research.att.com/~yehuda/presentations/drawing_large_graphs.ppt

Jurafsky, D. & Martin, J.H. (2000). Speech and Language Processing, Prentice Hall

Krenn B. (2000). Distributional and Linguistic Implications of Collocation Identification, Proc. Collocations Workshop, DGfS Conference, Marburg

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), pp. 211-240

Miller, G.A. (1995). WordNet: An on-line lexical database, Special Issue of International Journal of Lexicography, 3

Roget, P.M. (1946). Roget's International Thesaurus. Thomas Y. Cromwell, New York.

de Saussure, Ferdinand (1916). Cours de Linguistique Générale, Payot, Paris

Shannon, C.E. (1948). A Mathematical Theory of Communication, The Bell Systems Technical Journal, Vol. 27, pp.379-423, 623-656

Steyvers, M. & Tenenbaum, J.B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, subject to appear in Cognitive Science, 29-1

Tollis, I.G., Di Battista, G., Eades, P. & Tamassia, R. (1999). Graph Drawing: Algorithms for the Visualization of Graphs, Prentice Hall

Wendemuth, A.. (2004). Grundlagen der stochastischen Sprachverarbeitung, Oldenbourg Wissenschaftsverlag GmbH

Witschel, H. F. (2004). Terminologie-Extraktion: Möglichkeiten der Kombination statistischer und musterbasierter Verfahren, Content and Communication: Terminology, Language Resources and Semantic Interoperability, Ergon Verlag, Würzburg

Zipf, G.K. (1965). Human Behaviour and the Principle of Least Effort, Addison-Wesley Press