Project Deliverable Report

Deliverable nr D4.2 – Positioning service version 1

Work Package 4
Task 4.1 and 4.2
Date of delivery Contractual: 01-11-2009 Actual: 15-01-2010
Code name D4.2 Version: 1.0 Draft □ Final □
Type of deliverable Report
Security (distribution level) Public
Contributors Gaston Burek, Dale Gerdemann, Bernhard Hoisl, Robert Koblischke, Isobel Braidman, Alisdair Smithies, Debra Haley, Fridolin Wild, Petya Osenova, Kiril Simov, Adriana Berlanga, Christoph Maurerhofer
Authors (Partner) UTU, WUW, UNIMAN, OU UK, IPP-BAS, OUNL, BIT-MEDIA
Contact Person Dale Gerdemann, UTU
WP/Task responsible UTU
EC Project Officer Ms. M. Csap

Abstract (for dissemination) This report describes the implementation of version 1 of WP4 positioning services. The implementation is based on the design described in LTfLL D4.1 report and guided by tasks 4.1 and 4.2 scenarios. In addition this report discusses the theoretical basis for the implementation, the benefits for learners and stakeholders and the process for generating the data required to test, configure and initialise the services. Moreover, the report provides a technical specification for verifying the effectiveness of implemented language technologies and a description of how the positioning services can be integrated with other LTfLL services are also provided.

Keywords List Positioning services, knowledge rich approach, knowledge poor approach, conceptual development, LSA

LTfLL Project Coordination at: Open University of the Netherlands
Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands
Tel: +31 45 5762624 – Fax: +31 45 5762800
### Table of Contents

Executive Summary ........................................................................................................................................ 4  
Positioning service (task 4.1).................................................................................................................. 4  
  Knowledge poor approach .................................................................................................................. 4  
  Knowledge rich approach ................................................................................................................ 5  
Conceptual development monitoring service (task 4.2)........................................................................... 6  
1 Introduction ............................................................................................................................................... 7  
  1.1 Aims and background of WP4 positioning services ............................................................... 7  
  1.2 How the cultural and pedagogical theories are implemented by WP4 services ............... 8  
    1.2.1 Theoretical basis for positioning (task 4.1) ...................................................................... 8  
    1.2.2 Theoretical basis for monitoring conceptual development (task 4.2) ......................... 9  
    1.2.2.1 The self directed learner .......................................................................................... 9  
    1.2.2.2 Provision of feedback ............................................................................................... 10  
    1.2.2.3 Basis for the scenario - emerging group model ...................................................... 10  
  1.3 Benefits for learners and other stakeholders .............................................................................. 11  
    1.3.1 Benefits derived from task 4.1 scenario ........................................................................ 11  
    1.3.2 Benefits derived from task 4.2 scenario ........................................................................ 12  
2 Implementation of the positioning service (task 4.1)............................................................................ 12  
  2.1 Knowledge poor approach ........................................................................................................... 12  
  2.2 Knowledge rich approach ......................................................................................................... 14  
  2.3 Data requirements and positioning service initialisation ....................................................... 15  
  2.4 Assessing the learner position ................................................................................................... 16  
  2.5 Limitations of the current implementation .............................................................................. 17  
  2.6 Positioning service functionalities as described in the scenario ............................................ 17  
3 Implementation of the service for measuring conceptual development (task 4.2) ................................ 19  
  3.1 LSA ................................................................................................................................................ 19  
    3.1.1 Limitations on the use of LSA ...................................................................................... 20  
    3.1.2 Data requirements for the monitoring conceptual development service .................. 20  
  3.2 Network Analysis (NA) & Social Network Analysis (SNA) .................................................... 21  
  3.3 Fusion: Combining LSA and NA into Meaningful Interaction Analysis (MIA) .................. 24  
  3.4 The application: CONSPECT .................................................................................................. 26  
    3.4.1 Description ....................................................................................................................... 26  
    3.4.2 Required input ................................................................................................................ 29  
    3.4.3 Limitations of current implementation .......................................................................... 29  
  3.5 Usage example: learning about cancer ..................................................................................... 30  
    3.5.1 Interpreting Conceptograms ......................................................................................... 30  
4 How WP4 services relate to other services ...................................................................................... 34
4.1 Integration within WP2 .............................................................. 34
4.2 Integration within WP3 (threads) ............................................. 35
5. Verification and evaluation plans ........................................... 36
  5.1 For language technologies (in task 4.1) ............................... 36
  5.2 For CONSPECT (task 4.2) .................................................. 37
6. Conclusions and future work ............................................... 38
References .................................................................................. 40
Appendix 1. User interface ....................................................... 43
Appendix 2. Text management system ....................................... 48
Executive Summary

The WP4 services consist of a positioning service (task 4.1) implementing two approaches i.e. knowledge poor and knowledge rich approaches, and a conceptual development monitoring service (task 4.2). This report describes the implementation of the version 1 of these services.

Positioning service (task 4.1)

Knowledge poor approach

Version 1 of the knowledge poor approach implementation supports the positioning of the learner by means of learner language usage (quantitative and qualitative). Texts qualifying as instructional materials are used to initialise the service. In this context, quantitative analysis of language usage involves the use of quantitative information (e.g. token counts of distinctive phrases) in order to provide a measure of fit of learner language within the language used by a relevant Community of Practice (CoP). CoP language is reflected by instructional materials produced within the specific community. Qualitative analysis involves the identification of CoP distinctive phrases within learner texts. All qualitative and quantitative analysis outputs are intended as a source of information about patterns of language usage that may not be evident to the human reader. Positioning service users should interpret qualitative and quantitative outputs on the basis of additional knowledge about the learner that might be available in individual cases.

Work to date: In Burek & Gerdemann (2009), we have adapted and extended state-of-the-art language technologies, e.g. suffix array analysis (Yamamoto & Church, 2001) and statistical methods for non-parametric testing (permutation test) with the objective of developing a methodological framework for positioning. Based on these results, we have implemented and initialised a first version of the knowledge poor approach for positioning.

Methodology: The knowledge poor approach implements an extended LSA-based quantitative analysis and a qualitative analysis of learner texts. More specifically, the quantitative analysis measures the cosine distance between learner texts and a corpus of instructional materials covering concepts within relevant areas of expertise. The output of this analysis is a grade in a chosen scale for each learner text submitted for analysis. To conduct both the quantitative and the qualitative analyses the service requires a reference corpus. The service uses clustering and non-parametric statistical tests to suggest to users what prototypical texts should be used to build the reference corpus. Configuration of the positioning service and its associated text management service can be done by using web-interfaces.

Verification: The ongoing verification of technologies implemented by the knowledge poor approach compares training data results obtained with the traditional bag of words configuration against the alternative configuration that uses bag of phrases. Cross validation is then used over the chosen configuration.

Conclusions and future work: We have implemented and configured version 1 of the positioning services using the knowledge poor approach for a number of areas of
expertise in the computing science domain. Version 1 of the knowledge poor approach positioning services will be validated from the perspective of the user. Version 2 will be integrated with the knowledge rich approach in order to build a positioning service that can provide an enriched and more useful quantitative and qualitative feedback.

**Knowledge rich approach**

The knowledge rich approach supports the positioning of the learner by means of conceptual analysis of learner texts. The target position of the learner is defined by the means of a curriculum which in our case is understood in a broader sense - a curriculum for a course, job description or a questionnaire. The requirements, defined by one of these means, are grouped into topics which consist of two parts - textual description and conceptual description. The textual description is provided initially by tutors or admission panels. The conceptual description is constructed semi-automatically by the tutor. It consists of a set of concepts from an ontology that cover the semantic content of the text description. The learner will be presented with questions intended to cover these concepts. The degree to which the learner expresses knowledge of them (as measured by a kind of sentiment analysis) demonstrates his self-assessment of his knowledge concerning the given concept, which in turn provides indirect evidence of his actual knowledge of the concept.

The conceptual information, extracted from the learner texts, is compared to the conceptual description of the topics in the curriculum. A metric is defined on the basis of their overlap. The result is provided to the tutor. He does the final positioning of the learner, taking into account: the positioning suggested by the service, the actual learning texts and any other available information about the learner.

**Work to date:** In version 1 of the knowledge rich approach implementation, we have worked on the following tasks: (1) Improving the concept annotation pipe on the basis of the data made available by Bitmedia; (2) Extension of the ontology in order to provide support for better performing of positioning of the learner; (3) Translating the German data into English; (4) Extension of Bitmedia data with answers by Bulgarian students in IT, and generation of similar questions on the basis of the ontology, following Bitmedia patterns.

**Methodology:** In our work, we assume that the positioning of the learner is done on the basis of a set of pre-defined requirements, specified by a user who initiates the positioning - for us the users will be mainly the tutors, although in practice the user can also be a group of people (e.g. admission panel). The knowledge-rich method will rely on semi-automatic support to the tutors in their task. The tutor defines the concepts that need to be discussed by the examined person. Our tool will extract the mentioned concepts in the examinee’s test and will compare them to the pre-defined ones. This step is done automatically.

**Verification:** The verification of the technologies implemented by the knowledge rich approach is done by means of the same procedure used in the knowledge poor approach.

**Conclusion and Future Work:** For the first version of the knowledge rich service for positioning of the learner, we have implemented an appropriate ontology, lexicons in English and Bulgarian mapped to the ontology, and an annotation grammar. The extracted concepts from the learner texts are compared with the concept descriptions
included in the curriculum. Evidence of the conceptual coverage of the learner texts in the form of concepts lexicalisation is presented to the tutor who is responsible for the final positioning of the learner and for the recommendation towards further studies. For the second version, we will include sentiment analysis determining the level of self-assessed knowledge for each concept, extracted from the learner texts. There will be also a facility for navigation within the learner texts on the basis of the concept annotation. Better measures for the concept coverage of the learner texts will be applied.

**Conceptual development monitoring service (task 4.2)**

The conceptual development monitoring service provides learners a way to compare their conceptual development against different reference models, so they recognize the limits of their expertise. These models are (semi) automatically generated from learning materials and learner text inputs using Latent Semantic Analysis.

**Work to date:** WP4.2 has produced a service implementation that takes RSS feeds from learner blogs and produces concept visualizations, called “conceptograms”, mapped against a background corpus of domain content. For validation of V1 of the service with medical students, we are using PubMed abstracts for the domain knowledge but plan to add other sources. The service displays conceptograms from one source as well as conceptograms that are a combination of two others. The combined conceptogram indicates the differences between its constituent conceptograms.

**Methodology:** The technical development of the service uses LSA, trained by both a generic language and domain-specific corpus, so that texts from individual learners or from groups can be folded into a representative LSA space for analysis. The manner in which learners relate concepts together will be demonstrated by their degree of clustering and used as the basis for producing a visual representation. We are using a technique we named Meaningful Interaction Analysis (MIA). MIA takes the output from a standard LSA semantic space and displays it using ideas and tools from Network analysis, or graph theory. We use a force-directed lay-out algorithm to display the resultant graph.

**Verification:** We will evaluate the service accuracy in two areas. The first area is the agreement between the semantic similarity of concepts calculated from a latent semantic space and test participants. The second area is the extent to which test participants agree with the concepts extracted from a text by the service. We have devised three experiments to test these two areas of accuracy.

**Conclusion and Future Work:** The service as currently implemented (Version 1) provides a snapshot of the conceptual development of a learner. We are investigating other ways to display the snapshot, e.g. lists of concepts.
1 Introduction

1.1 Aims and background of WP4 positioning services

This report describes the implementation of version 1 of the WP4 services, which includes the positioning service and the conceptual development monitoring service (concept inspection service: CONSPECT).

The positioning service provides support to the education provider’s staff and learners in understanding their learning history and current competences. Based on the analysis of learner texts, the service assesses the learner’s position on the basis of the learner’s language usage (knowledge poor) and concepts covered in the learner’s texts (knowledge rich) to inform and support users e.g. tutors in the task of recommending suitable learning material.

The CONSPECT service is designed to provide information about conceptual development for both individual learners and their tutors. It uses text output from learners, both as individuals and as groups, to establish a visual model of the concepts learners use and enables learners to be compared with one another. It will be available to individual learners and to their tutors to provide a basis for monitoring an individual’s progress, the development of a group of learners, and for relating the progress of an individual to that of his/her peers.

The current implementation of the services is based on the design presented in LTfLL deliverable D4.1 (D.4.1) and the LTfLL Consortium’s Approach to Integration (CAI)—additional report that forms the foundation for pedagogical context, specific to the deliverable, as exemplified by the scenarios. The language technologies implemented by WP4 services, already discussed in D4.1, go beyond the state of the art and are based on well established theoretical frameworks that support social and individual learning and are aimed to capture learners’ knowledge at two different levels, i.e. linguistic and conceptual. While the positioning service bases its analysis on linguistic information derived from learner and education materials texts, CONSPECT analysis is based on concepts inferred from those texts. The validation of the current implementation will guide the further development of language technologies and usability features to be implemented in version 2 of the services.

This report is therefore structured according to the integrated approach and will consequently describe the theoretical basis, pedagogical context, development of services, and validation as a basis for the next version of the services. The rest of this section provides background. It reminds the reader about the aims of the WP4 services and describes how cultural, pedagogical and linguistic theories have informed the services’ implementation in accordance to the scenarios developed in WP3. Section 2
describes the implementation of the positioning service and section 3 describes the implementation of the CONSPECT service. Section 4 describes how WP4 services could be integrated with other LTfLL services by means of threads. Section 5 discusses validation plans and section 6 draws conclusions and describes future work.

1.2 How the cultural and pedagogical theories are implemented by WP4 services

1.2.1 Theoretical basis for positioning (task 4.1)

As explained in D4.1, the positioning service has two main theoretical perspectives, i.e. cognitive and social. On the one hand, the cognitive perspective assumes that memory and knowledge play the most important roles during learning. On the other hand, the social perspective assumes that the learners interact with their learning environment and are influenced by it. In this respect, the major features of the social perspective are social engagement in communal activity and the identities, language, and cultural artifacts of the social groups in which individuals learn by interaction. Within this perspective, learning is a collaborative activity. Social cultural interactions generate shared tacit knowledge as well as linguistic patterns of usage (speech genres; Bakhtin, 1986).

The service incorporates the social perspective by implementing a qualitative analysis that compares a learner’s language usage to language usage within specific CoP. An example of CoP-specific terminology is provided by the terms “Latent Semantic Analysis” and “Latent Semantic Indexing”. The two terms mean the same thing but the latter term is more characteristic of the Information Retrieval CoP. Other related terms that are sometimes used in more specific CoPs are “Latent Semantic Space” and “Latent Semantic Network”. The details of which terms are appropriate in which contexts are complex and subtle. Incorrect terminology use is detrimental to communication within a CoP and is indicative of non-integration into the CoP. Given that these usage patterns are difficult for humans to observe, language technologies, and particularly an approach based on a bag-of-phrases model, can be usefully applied to provide such feedback.

In addition, the positioning service incorporates the cognitive perspective by means of evaluating which concepts have been covered in learner texts. The objective of this analysis is to evaluate the learner texts’ coverage of tacit and shared knowledge formalized as an ontology.

The main prerequisites for the application of the knowledge rich approach are the domain ontology and the lexicon(s). However, due to the pace of rapidly developing technologies, the need for such resources turns out to be also one of the shortcomings of the approach. In practice, these resources tend to lag in their development behind the development of the domain itself. Consequently, it becomes necessary to ensure constant enrichment of the ontology and the lexicons in order to cover new concepts and lexicalizations. Such an enrichment is not meant as chaotic, but rather governed by a particular task. We envisage to provide a service which supports semi-automatic ontology and lexicon enrichment. The primary source for new lexicalisations comes from the
phrase extraction component of the knowledge poor approach. The tutor will have the possibility to examine such phrases and to include them as new lexicalisations of existing concepts, or to create new concepts with appropriate lexicalisations.

1.2.2 Theoretical basis for monitoring conceptual development (task 4.2)

As described in D4.1 and in D3.2, the learning context is that of lifelong learning in a workplace learning environment. Problem Based Learning (PBL) and Enquiry Based Learning (EBL) are used to exemplify some of the group interactions, which are characteristic of this milieu (Tipping et al., 1995). Social learning theories have particular relevance to these approaches to learning, as they depend upon an interaction between the individual learner and the PBL or EBL group, who may develop towards a “community of practice” as described by Lave & Wenger (1991), where learning and understanding proceed collaboratively. Such interactions also develop in communities of online learners, in which there is an interplay between cognitive development, social interactions and tutor presence (Garrison et al., 2000), as used to promote reflective learning in medical students (Braidman et al., 2008). The continuum of personal knowledge building to a generally held understanding has been described in detail by Stahl (2006). We have discussed in detail the relevance of Stahl’s learning cycles to task 4.2 (D4.1), and this has also been analysed in tasks 5.2 and 6.2, but it is important to note here the basis it provides for understanding the learner’s development in a PBL (or EBL) learning approach. We have also previously discussed the theoretical basis on which expertise is acquired. In the context of this work package, where the learner can develop a self directed role, the Model of Domain Learning, MDL (Alexander, 2003), is of relevance as it recognises that with increasing expertise, domain knowledge undergoes both quantitative and qualitative changes, learning strategies are developed that are related to depth of domain knowledge and interest invested by the learner varies in both general and specific contexts. Schmidt et al. (1990) describe the application and association of concepts to clinical reasoning processes for problem solving purposes in the medical domain. It requires the restructuring of knowledge and linking between concepts as described by Alexander (2003). This is the key to the educational context and scenarios described by this work package, where learners are required to link concepts, derived from the evidence included in a medical PBL case, in order to understand the implications of certain clinical conditions, their investigation and their treatment. Attainment of this degree of expertise depends both on progress as a self directed learner and the learner’s interactions with his/her peers on group discussion, both face to face and online.

1.2.2.1 The self directed learner

The importance of self directed learning or self regulated learning (Puustinen & Pulkinnen, 2001) has been discussed in detail in task 5.2. Here the learner adopts different stages of a self directed learning cycle (Butler & Winne, 1995), namely task definition, where the learner identifies the specific learning task, followed by goal setting and planning, in the context of prior knowledge, enactment, of the plan and adaptation, where the learner adapts his or her cognitive structure in the light of the past learning
event. In the medical learning context, this approach is encapsulated by that of Schön’s reflective learning cycle (1987), where the ability to think critically is crucial for an individual to understand limitations of knowledge, competencies, skills, attitudes and behavior and to modify actions and learning plans. This approach underpins the learning context of task 4.2, where provision of feedback enables the learner to understand his/her progress in linking concepts in order to address clinical problems, to identify limitations and to formulate learning plans that address these issues, which can be modified further. In the PBL/EBL learning context, learner development is a result of interaction between understanding gained as an individual and as a member of a group. This interplay is important in developing the means of providing feedback in task 4.2.

1.2.2.2 Provision of feedback

In order for feedback to be fully effective it should be appropriate to the learner’s zone of proximal development (Vygotsky, 1978). For the educational context of task 4.2, this must reflect the common understanding achieved through interaction with the group of peer learners but must also indicate the knowledge held in common by the community of practitioners with expertise. In D4.1, we discussed the basis for reference models which would achieve this (Berlanga et al, 2009) and identified archetypical reference models (expert and state-of-the-art information e.g. scientific literature), pre-defined reference models, which are curriculum based (e.g. course material, tutor notes, relevant reading materials, etc.), and emerging group models that represent concepts and the relations between them as used most often by a learner or practitioner group (e.g. peers, participants, co-workers, etc.). We also presented evidence that although pre-defined reference models could be important in driving forward an individual’s understanding of a particular domain, the emerging group model was closer to an individual’s zone of proximal development and would be likely to provide more support near the beginning of a learning trajectory.

The aim of the task 4.2 service is to provide tutors with evidence on which to base facilitative feedback (Shute, 2008). Feedback types have been defined in terms of complexity by Shute (2008) and are discussed in some detail in the LTfLL WP5 report (D5.1). The service will provide feedback relating to attribute isolation of a target concept, and will highlight learners’ misconceptions.

1.2.2.3 Basis for the scenario - emerging group model

The emerging group model is important to the reality of the PBL/EBL learning context. The notion of the self directed learner who, through interaction with a peer group, develops understanding of important domain concepts and, thereby, analyses and solves problems, is a challenge to novice learners. Nevertheless, it equips them with the skills and competencies that will support them throughout lifelong learning. The use of the emerging group model will enable tutors to understand the relationship of the individual learner to the rest of his/her peer learning group and to analyse the range of development in the group as a whole. Learners who are either ahead of or behind the rest of their peers can therefore be identified in order to provide them with the appropriate support for their further development. Analysis of problems and issues in a PBL/EBL context usually
occurs in face to face discussions between peers, but online discussions between learners are now becoming part of the PBL process (Woltering et al., 2009). Text output from individual learners and their group interactions are therefore available for analysis (Braidman et al., 2008) and could be used to generate the emergent reference model using text from online discussions concerning specific PBL cases.

**1.3 Benefits for learners and other stakeholders**

As explained in section 1.1, the positioning service implementation has been guided by the corresponding tasks scenarios 4.1 and 4.2 (see D3.2). Both scenarios describe stakeholders' roles, users of the services and functionalities to be implemented by such services.

**1.3.1 Benefits derived from task 4.1 scenario**

Task 4.1 scenario explains that the Austrian Public Employment Service (AMS) matches unemployed individuals with job openings. If candidates lack specific skills for typical job offers, the AMS offers them qualification opportunities (e.g. in the IT domain). The AMS defines the skills requirements for each candidate and delegates training activities to education providers (e.g. Bitmedia) that have to offer efficient learning environments for those candidates to meet the AMS requirements and furthermore to acquire the non-well defined, socially important skills that allow for integration into the relevant CoPs. The use of the positioning service will enable the tutors (employed by the education providers) and learner candidates to establish a more specific definition of the required training units and corresponding learning methods. The provision of detailed and optimised hints for the learning path enables the learner to save time during the learning curve. This implies also that learners and tutors will be more satisfied and motivated. Using the enhanced positioning procedure will enable education providers to develop individualised and cost reduced training for the unemployed individuals. From the point of view of the AMS, the enhanced positioning process will reduce the overall cost per individual (for the required training and learning).

Unemployed individuals are a non-homogeneous group in terms of their acquired experience and skills. Conventional training approaches, therefore, do not cover their pedagogic needs. Learners need individualised training to meet the learning requirements as specified by the AMS and socially required by the relevant CoP. The positioning service provides on demand feedback to the learner and tutor, with only minor increase in workload.

Based on the feedback derived from the analysis of the learner's free text, both semi-automatic approaches implemented by the positioning service (i.e. knowledge poor and knowledge rich) support the tutor in providing the appropriate learning materials and learning methods. The individualised learning path, which is more focused on the required information and the appropriate learning methods, will be expected to establish a learning environment which fits the expectations of the learner and increases retention. This enhancement of the learning settings is also expected to increase tutors satisfaction,
because they are able to focus their engagement on supporting the individuals during the learning process.

### 1.3.2 Benefits derived from task 4.2 scenario

The main stakeholders of CONSPECT service are learners and tutors. The service has been designed to support tutors, who administer their populations of learners, and learners, who are able to contribute materials, view their progress and share their concept maps with others. Learners will be able to review their progress, based on the evidence they submit, over time. The service is intended to provide an indication of conceptual areas where there are opportunities to enhance development through formative feedback. Formative feedback is provided by tutors, to whom the service identifies the learners’ conceptual gaps. The key benefit of the service is its ability to use evidence to diagnose potential development needs in a timely manner. The service has the potential to save tutors time in the task of assessing individuals for their formative feedback requirements. Students will benefit from feedback interventions more appropriate and timely to their learning needs.

The pedagogic benefits of the CONSPECT service are rooted in expertise development. By automating the analysis of a learner’s coverage of concepts in their written materials and providing this as a visual cue to the learner, they are able to see the extent to which they have provided evidence for each of the concepts that make up a specific domain of knowledge. The learner can then take self-directed actions to address a knowledge area that is identified as deficient, for example by contributing further materials, or they may choose to share their concept map with peers or their tutor. When learners choose to share concept maps with their tutors, the tutors are able to see under-represented areas of knowledge, for which they can identify a course of remedial action that combines suitable resources and learning activities. The tutor may also choose to alter his or her learning activities for the group of learners, based on a concept map that represents a group reference model.

### 2 Implementation of the positioning service (task 4.1)

The positioning service positions learner texts by means of (1) a knowledge poor approach and (2) a knowledge rich approach. Results from both approaches are independent of one another. The knowledge poor approach implements a quantitative and a qualitative analysis of learner texts. Results from those analyses reflect the learner’s integration in a community of practice (CoP) by looking at how representative the language used by the learner is of the language used within a particular CoP. The results from the knowledge rich analysis of the learner’s text reflect coverage of conceptual knowledge associated with the specified CoP. The main users of the service can be admission panels, tutors and learners themselves.

#### 2.1 Knowledge poor approach

The knowledge poor quantitative analysis measures the cosine distance between learner texts and a corpus of instructional materials covering concepts within relevant areas of
expertise of a CoP. The output of this analysis is a grade for each learner text. The relevant concepts are specified by a learning goal (or job description) that has an associated list of concepts that the learner is required to know.

The design of the service is based on the idea that each CoP develops its own “speech genre” (Bakhtin, 1986) which may be reflected in the instructional materials generated within the CoP. As explained in Burek & Gerdemann (2009), the Latent Semantic (LSA) bag-of-words approach presents many shortcomings when applied to the positioning task. To solve this problem in addition to the quantitative analysis the service implements a qualitative analysis that is used to improve LSA by means of incorporating into the analysis information about learners’ language usage. The quantitative analysis represents texts using a bag of distinctive phrases instead of a bag of words. The qualitative analysis involves the scoring of phrases extracted from learner texts according to distinctive features of their usage to describe competences within a CoP. In this context the “distinctiveness” of a phrase is a statistical measure of how characteristic of document quality the phrase is by comparing its frequency of occurrence in high and low quality texts as graded by experts. In addition, the qualitative analysis output provides justification for the quantitative analysis giving the users (e.g. learner and tutor) some confidence that the score is based on the analysis of the learners’ written phrases and not simply based on word frequency. Users can inspect the scored phrases visually.

To understand the meaning of distinctiveness in more detail, let us consider the following example. The phrase “linear system is a set of” occurring a total of 10 times within documents categorised as linear algebra. The phrase occurs 5 times within documents graded one, 3 times within documents graded two and 2 times within documents graded three. If the service gives 1 point for every text graded one, 0.8 points for every good text graded two and 0 points for every text graded three the phrase receives a first score of 7.4. To evaluate how distinctive the score 7.4 is, the service simulates 1000 scores by randomly assigning a grade to the 10 documents. Then, the service calculates a new score by counting how many of these 1000 scores have a value of 7.4 or more. This last score is used to characterise the distinctiveness of the phrase under analysis. A similar Monte Carlo approach is taken to determine the distinctiveness of phrases which are characteristic of poor texts, so that the test for distinctiveness is a two-tailed test.

To calculate the statistical measure of distinctiveness, the service needs fine tuning for each set of the relevant concepts by means of training using representative texts of language usage to describe those concepts. To conduct both the quantitative and the qualitative analyses, the service needs a reference corpus containing the representative texts. To build that corpus the service analyses instructional text materials that are made available by experts, e.g. by uploading the materials in a repository accessible to the service (see appendix 1 and 2). Each new text induces a closeness ranking to all texts already included in the manually built initial reference corpus. A non-parametric permutation test is used by the service to measure the goodness of this ranking. The service uses the texts in the initial reference corpus as prototypes to suggest what texts from the instructional materials should be added to it. Based on this measure, the user
decides whether or not to add the new text. The reference corpus is thus built incrementally every time new instructional materials are made available in the service repository. The initial reference corpus is built manually using texts corresponding to learners’ answers for questions covering a different set of concepts from an existing ontology. Experts grade these answers with a suitable grading scale and upload (into the service) each question together with the corresponding answers and set of concepts. The service only incorporates answers into the corpus which are graded with the highest grade (by default) or with a lower grade if specified by experts.

Every time a new set of concepts becomes relevant that is not yet covered by the reference corpus, experts can generate new questions covering those concepts and then upload into the service each question together with its corresponding set of concepts. Answers to those questions will later be graded by experts and manually added to the reference corpus. In addition, the second version of the service will incorporate the analysis of online discussion forums. Each forum will discuss a set of concepts. Learner contributions to the forums will be annotated with the relevant concepts and graded by experts. High quality forum entries (to be implemented in version 2 of service) can also be added to the reference corpus as instructional material.

2.2 Knowledge rich approach

This approach covers concept extraction from texts and sentiment analysis related to the concepts. While the output of the knowledge poor quantitative analysis provides information about correctness in the style of learner texts, it does not provide any information about their conceptual coverage. Functionalities implemented by the knowledge rich approach compare to those of task 4.2. But, while 4.1 knowledge rich analyses learner texts with the objective of identifying concepts that are formalised in an available ontology, task 4.2 identifies concepts from learner texts by means of clustering without external knowledge about those concepts and their lexicalisations.

For measuring conceptual coverage or completeness of learner texts, the service uses an ontology to count how many of the concepts specified by the curriculum are covered in the learner texts and outputs a value that reflects the percentage of the relevant concepts covered. In version 1, the service displays the existing ontology and a list of concepts extracted from the learner texts. The extracted concepts are represented by the phrases resulting from the knowledge poor qualitative analysis. The tutor uses the ontology in order to semi-automatically define the relevant concepts. Then the concepts extracted from the learner texts are compared to these specified by the curriculum. In this way, the service provides evidence of the conceptual coverage of the learner texts.

In version 2, language technology for sentiment analysis will be implemented to identify learners' assessment of their own knowledge about the extracted concepts. The sentiment value for a concept is positive if the evidence in the text suggests that the learner thinks that the concept is known and the value is negative if the evidence suggests that the learner thinks the concept is not known. Normally, the service considers a mentioning of a concept as evidence that the concept has been covered. Negative sentiment toward the
concept suggests, however, that the concept was not covered, even though the concept is mentioned. Positive sentiment toward the concept indicates only self assessment, which the tutor must evaluate in the light of other evidence. The usage of ontology as a formalisation of the conceptual content of a curriculum provides a possibility for more exact measurement of the concepts in the domain known by the learner, their realisation in the learner’s texts by different phrases and ability to work in multilingual settings - curriculum prepared in one language and learner’s texts in different languages.

Current implementation of the knowledge rich approach produces two analyses from the learner texts. The first analysis supports users (e.g. tutors) in identifying concepts from a domain ontology in a curriculum (e.g. questionnaire) and is implemented as a RESTful web-service including a language pipe with a tokenizer, a POS tagger, a concept annotation grammar, a co-referential relation module. The input for this analysis is pure text, and the output is the same text, but annotated with the appropriate concepts from the ontology. The annotated texts can be used further as the source for extraction of conceptual information needed for measuring conceptual overlapping and in addition as visualisation of the actual usages of a given concept.

The second analysis supports users in measuring the conceptual overlap between learner texts and the curriculum. This analysis takes as input two sets of concepts - one corresponding to the concepts covered by the curriculum, and one with the extracted concepts from the learner texts. It returns the percentage of the overlapping concepts between the two sets and also three distributions: (1) the common set of concepts, (2) the set of concepts not covered by the learner texts, and (3) concepts that are covered in the learner texts, but are not covered by the curriculum.

2.3 Data requirements and positioning service initialisation

A prerequisite for the service is the availability of learning materials, questionnaires and learners texts (answers), an ontology in the specific domain and lexicons mapped to the ontology. Tutors employed by the education provider are responsible for the initial configuration of the positioning service. Based on the skill requirements defined by AMS and practices and expertise associated CoPs, tutors collect the available instructional materials and existing questionnaires for the preparation of a new questionnaire. The new questionnaire is built by reusing available questions or creating new ones that cover concepts associated to learning requirements and that can be mapped to available training materials. Then, answers to the questionnaire are collected from ten or more individuals. These answers are graded by at least two tutors who will upload the whole data set using the graphical interface described in Appendix 1, with the objective of initialising the process of building the reference corpus. Subsequent proposed additions to the corpus will be rated based on the ranking they induce on the documents existing in the corpus, and that the tutors can take this rating into account when deciding whether or not to extend the corpus.

After the initialisation, the positioning service for this specific questionnaire associated course is prepared for the first use with the minimal amount of required materials. The
service will get better after each use, since well-answered questionnaires can be added to the corpus.

The knowledge rich approach uses the ontology for representing the conceptual knowledge within the domain. The lexicon contains terms for the concepts within the ontology. On the basis of the lexicon, a concept annotation grammar is created. It is used for recognition of the concepts within the learner texts. We support lexicons and annotation grammars in English and Bulgarian. These requirements can be defined as a curriculum, consisting of a set of topics and related learning materials. Each topic can be given a simple question or a short description. The learner writes answers to the questions or a short text on the topic. Each topic is then represented by a set of concepts from the ontology. These concepts are offered by the service after the topic descriptions have been processed. The tutor is provided with the possibility of editing the offered list. He or she is able to include more concepts from the ontology or also to exclude concepts. After editing, the selected concepts are stored as formal definition of the curriculum. The learner text is processed by the service. Then the concepts, mentioned in the text, are extracted. These concepts are sorted into several lists according to the contexts in which they are mentioned (this part will be implemented in the next phase). The extracted lists are compared to the list of concepts pre-defined by the tutor. Depending on the degree of overlap, the learner is positioned on purely technical terms on the scale between good and bad. For example, if the overlapping is 80%, the student would be considered good. However, this estimation does not replace the tutor's role, but just provides useful additional information. The tutor (or another user) can see the extracted lists, the pre-defined list of concepts (from the curriculum) and the degree of their matching. Then, adding the qualitative perspective, the user can accept, modify or reject the suggested positioning. For this task, the ontology will need to be extended with additional concepts which cover areas of expertise associated to the vacancy as identified by the education provider (i.e. Bitmedia) and new relations between existing concepts (will) need to be established.

### 2.4 Assessing the learner position

Task 4.1 service supports the assessment of the learner position in the following way: A workshop (about four hours depending on the course) is scheduled. During the introduction, each 'learner' is sensitized to the relevance of providing detailed and long answers. Learners answer questions which ask about their level of knowledge concerning a certain relevant concept, and will participate in forum discussions for Version 2 of the service. The service compares the text generated by the learners with texts in the reference corpus by means of analyses implemented in the knowledge poor and knowledge rich approaches as described in the previous sections. Finally, after examining the output, tutors can decide what materials need to be studied by the learners, and what kind of support learners may need, e.g. help in improving writing skills or conceptual knowledge, etc. This decision concerning the learning path (materials and methods) for the 'learner' are discussed and finalized in the positioning task final session between 'tutor' and 'learner'. During this session, learners get an introduction to their individualized learning path.
2.5 Limitations of the current implementation

The knowledge poor approach requires the use of a reference corpus to be able to identify distinctive phrases. This requires tutors to manually upload such materials into the service on a regular basis. After discussing this with the education provider (Bitmedia), the conclusion was that uploading materials will not impose any significant workload to tutors because the uploading process does not require extensive evaluation of the materials. The corpus refinement procedure incorporated in the knowledge poor approach will automatically evaluate the materials and suggest what texts should be included in the reference corpus.

As mentioned above, every time a vacancy requires the education provider to cover a new area of expertise, the ontology used by the knowledge rich approach will have to be extended with additional concepts and new relations between existing concepts. This shortcoming is counterbalanced by the higher precision that this approach offers.

In version 2, the phrase extraction functionality implemented by the knowledge poor approach will be used to identify lexicalisation of the ontology and, in some cases, to partition the set of lexicalisations of a concept to split concepts in the ontology into sub-concepts or related concepts.

2.6 Positioning service functionalities as described in the scenario

The first prototype for the version 1 of the positioning service was built around task 4.1 scenario requirements. An overview of the services and data is provided in Figure 2.1 as a generic Service Usage Model (SUM). The business processes (blue boxes) are in the first row - classified in three phases (course definition, setup & initial configuration, and learning phase). Underneath each business process the individual user-facing and background services are described. Used data sources are displayed at the bottom of Figure 2.1.
The basic functionalities implemented so far encompass:

1. Course management - create, list, edit and delete course (or communities of practice) information
2. Questionnaire management - create, list, edit and delete questions or topics for each course
3. Answers management - create, list, edit, delete, grade and give feedback to answers for each question
4. Upload and manage reference material - upload, download and delete learning materials for each course
5. Automated and human feedback - provide automated and both qualitative (listing significant and quality-distinct phrases) and quantitative (grade estimates) feedback, to support both the answering student and the grading tutor with automated and live feedback as well as knowledge rich feedback consisting in concept identification and comparison.
6. Semantic space calculation and management - create, edit, build and train latent semantic spaces as a basis for quantitative feedback based on LSA as well as facilities to incorporate distinct phrase extraction software developed by this task.
7. User authentication - based on platform-independent OpenID accounts.

We implemented all of these services as widgets based on the Wookie engine, so we designed the interfaces for that and integrated it in a first step in a site laid out with HTML (see Appendix 1). We have also strived to keep the interface familiar across all these modules, as the basic tasks (create, edit, list, delete, etc.) are often recurring as visible above. Besides giving us the ability to relate documents/texts to user IDs, the user authentication also allows us to have different roles in the system such as the ones defined in our scenario. Login can be done by both local login with user and password credentials and also with an OpenID identifier which can be used throughout all WPs widgets as this will be the authentication method everyone will use by integrating all our services into one personal learning environment.

3 Implementation of the service for measuring conceptual development (task 4.2)

This section provides an overview on the approach developed within LTfLL to support the monitoring of conceptual development. It first mentions some of the limitations of Latent Semantic Analysis (LSA) and introduces Social Network Analysis (SNA). Then it shows how the output of LSA and concepts from SNA are merged into Meaningful Interaction Analysis (MIA) to create a visualisation, or conceptogram that allows for the analysis of conceptual interactions in latent semantic vector space representations. It concludes with examples of conceptograms and instructions to interpret them.

3.1 LSA

LSA is the theory that underlies much of LTfLL work. This is a statistical method for capturing meaning from text. A seminal paper (Landauer et al., 1998) gives a more formal definition: “Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text”. LSA was developed as an information retrieval technique in the late 1980s, when it was called Latent Semantic Indexing (LSI) (Deerwester et al., 1990). Later, the developers found that LSI could be useful to analyse text and created the term LSA to describe LSI when used for this specialised area.

LSA “induces global knowledge indirectly from local co-occurrence data in a large body of representative text” (Landauer & Dumais, 1997). The LSA technique is essentially a method for solving a huge set of simultaneous equations that represent terms in documents (Landauer, 2007 pp. 13-14). The mathematical underpinnings of LSA can be daunting to non-mathematicians (Hu et al., 2007 p. 407). They state “Why and how it works is a very deep mathematical/philosophical question …” and point the interested reader to Martin & Berry (2007). A technical description of how LSA works can be found in previous deliverables.
3.1.1 Limitations on the use of LSA

This section briefly mentions some of the criticisms about LSA that some researchers have made. LSA takes no account of word order. In the often cited example, “Mary loves John” is not the same as “John loves Mary”. LSA uses the so-called bag of word approach, which removes any word order information. This seemingly insurmountable problem turns out not to be fatal.

Another criticism about LSA was made by McCurley (2009), a researcher at Google, who reported that a Google team investigated LSA as a search technique in 2005. He went on to say that the results were bad prompting Google to abandon the idea of using LSA. However, the details of what McCurley said are very relevant to LTfLL. McCurley stated that the results for the tail-end of the web were “ok”. Documents in the so-called tail-end are “jargon rich” where the “term overlap means something”. As most of Google’s traffic does not involve tail-end searches, LSA was not appropriate for Google. The good news for this project is that the kinds of text we attempt to analyse are precisely those jargon rich documents for which McCurley reported “ok” results.

Finally, LSA demands a certain amount of calibration and the results are dependent on the calibration choices made by the developers of a particular LSA implementation. One of the factors that must be decided is the composition of the background corpus. An inappropriate background corpus cannot produce good results. A very simple example makes this point clear. Consider the terms Cuba, Jamaica, and Russia. If the background corpus comprised mostly tourism-related documents, then LSA would calculate that Cuba and Jamaica are more closely related than Cuba and Russia. However, if the background corpus comprised mostly political documents, then Cuba and Russia would be more closely related than Cuba and Jamaica. One of the WP4 partners has suggested an automatic technique for varying the composition of the corpus that he calls corpus refinement. Task 4.2 involves creating a corpus, which as currently implemented is done by an administrator - there is no automatic support for corpus refinement, though first experiments have been conducted.

3.1.2 Data requirements for the monitoring conceptual development service

An LSA process can be divided into the following steps: text pre-processing, weighting, calculation of the SVD, and closeness measurement plus utilisation. Each of the steps offers a variety of options to configure the actual analysis. How to calibrate, however, seems not to be a simple question, as many contradicting claims can be found in the literature concerning the adjustment of these influencing factors. Perfetti (1998), for example, argues for more reliable results with a larger corpus size (input documents). On the other hand, Deerwester et al. (1990) were able to successfully apply LSA to a corpus with only nine documents. Nakov (2000, 2001, 2003) reports the best results with a raw term frequency applied as local weighting scheme, whereas Dumais (1990) finds a
logarithmized term frequency suiting best. Dimensionality is seen quite different by authors. Dumais (1990) sticks to the magical number of 100, whereas Graesser et al. (1999) suggest the use of 100 to 300 dimensions. Nakov (2000) recommends the number of dimensions to vary from 50 to 1500. Conclusions on how to calibrate can hardly be drawn from these statements. Furthermore, these examples indicate that identifying the perfect calibration is complex and tightly coupled to the purpose the application serves. The various influencing parameters may even hinder users in calibrating LSA to achieve optimal results, sometimes even lowering performance below that of the (quicker) simple vector-space model.

Typically, proposals on how to create latent-semantic spaces stay either agnostic about the required size of a corpus while at the same time warning not to use too small ones (Quesada, 2007) or they recommend the use of bigger corpora with an unreflected ‘bigger is better’ assumption. Several authors report on positive results with relatively small corpora and encourage research on their composition: amongst others the inventors of latent-semantic analysis (LSA) provide in their disseminal paper (Deerwester et al., 1990) a very convincing example utilising only nine documents.

The underlying problem is a problem of creating a latent-semantic space big enough to cover the desired target domain and apt to separate the relevant senses, connotations, and meanings from the – in this field and for this application – irrelevant, but at the same time small enough to allow for computational efficiency. This is a question of validity: can convergent and discriminant validity be proven for small spaces.

As of today, it is not clear, how to create such a corpus efficiently and without extensive validation. However, several requirements can be uttered that constrain the possibilities for an application service as being created for task T4.2 monitoring conceptual development. Most notably, the availability of representative text material constrains efficiency: for the best, both background and domain-specific (community of practice specific) texts required should be openly available and the costs of gathering them should be reduced to a minimum. Typing in thousands of hand-written student essays, for example, would be a huge barrier to uptake of such a technology developed. Reliability is another important requirement: student writings investigated now should produce the same results when repeating their evaluation a month later. Spaces should provide stability, as otherwise the interpretation of the presented findings on conceptual development is endangered. Moreover, aggregating an emergent reference model over a multitude of individual cases becomes impossible, if the stability of the space cannot be guaranteed over time. Efficiency constrains uptake: it needs to be possible to calculate spaces from corpora on typical industry-sized machines, i.e. what is expected to be widespread standard technology towards the end of the project and in the long-term five to ten year perspective, the LTfLL project has been funded for.

3.2 Network Analysis (NA) & Social Network Analysis (SNA)

The raw output from an LSA process is difficult to interpret. The fields of Network Analysis and Social Network Analysis can provide guidance about how to display the output in a way that can be more easily understood, that is, as a graph. LSA graphs are not, strictly speaking, social networks. However, social networks can be thought of as a
A metaphor for LSA networks: social networks show the inter-relationships among (usually) people; LSA networks show the inter-relationships among words and documents.

The term SNA was coined in 1954, but precursors and research in this area date back to at least the latter 19th century. Graph theory, also known as network analysis, goes back to 1735 with Euler’s Seven Bridges of Königsberg problem. The basic idea of SNA is to provide a means to analyse actors and the relationships between them (Monge & Contractor, 2003). Actors thereby are often people, but can also be groups, media, tags, or any other acting entity that an analysis focuses on. Ties of different type and strength connect these actors. Together actors (also known as ‘nodes’ and ‘vertices’) and ties (also known as ‘edges’ and ‘links’) form a graph. Within this graph, certain structures can be identified and standard NA techniques can be applied to find them. Ongoing research provides a growing variety of instruments for investigating structure and structural properties in this graph as a proxy for deriving information about the relationships in the real world. It is of particular relevance to this work package as it may form a paradigm for the emerging reference model, discussed in 1.4.4.

Social network data has been displayed with the help of sociograms, (see Figure 3.1) which are visualisations that use an optimised lay-out algorithm to project a complex graph structure onto a 2D display in a way that it most closely resembles the actual structure (see Fruchterman-Reingold (1991) for force-directed lay-out algorithms.) Understandably, this projection becomes harder to interpret and – even more problematic – less precise and reliable, when the underlying graph structure is big and complex, which is often the case with social networks.

Figure 3.1 Sociogram.
As a way around this misinterpretation pitfall, network analysis provides a variety of measures that are both visual and non-visual in nature to more accurately investigate the nature of the graph structure. The next paragraphs explain the terms in Figure 3.2. See Brandes & Erlebach (2005) for more information about these, and other, measures.

Figure 3.2 Different network measures illustrated.

To mention just a few of them: on the level of an individual node, the degree centrality of each node can be calculated (see Figure 3.2, top left: the node size is scaled by its degree centrality). The degree centrality counts the number of in- or outgoing connections of an individual node. In case of undirected networks, the degree centrality counts the number of connections (regardless of direction). Nodes with a high (in or out) degree are more central in the network under investigation, e.g., more communication passes though them. Nodes with a low degree centrality are less connected than the central nodes.

Betweenness is a figure that counts how often the node is an intermediary between two not directly connected nodes.

Closeness is an indicator measuring how close the person is to all other members of the network; it is measured as inverse sum of the shortest path lengths (geodesics) between individual nodes and all others.

Network or component level indicators are, e.g., the network’s (component’s) density reflecting how connected a group of networks is. It is commonly measured by calculating
the number of actually realised connections divided by the number of possible connections.

In the example above, there are 29 realised edges, and $18 \times (18-1) / 2 = 153$ possible connections. Therefore, the density of this network is with .19 relatively high – as the network density typically tends to decrease with growing network size.

Other indicators such as centralisation reflect how hierarchical a network is. The clustering coefficient gives a measure of how structured a network is (how ‘cliquish’ a network is).

### 3.3 Fusion: Combining LSA and NA into Meaningful Interaction Analysis (MIA)

The graphs created with latent semantic analysis form a complex network structure expressing a manifold of relations between the core actors (terms and documents). In other words, a multitude of differently weighted edges between the nodes is what the resulting spaces and matrices express. Naturally, this network can be investigated with network analysis – a field of research made popular not least by social network analysis and its achievements in especially the last century. We call this latent semantic network analysis meaningful interaction analysis (MIA).

Since every vector in an LSA network has a cosine with respect to every other, the resulting network is in its unfiltered form a completely realised network with the density equaling or very close to 1. Figure 3.3 shows an example.

![Figure 3.3 Resulting (unfiltered) network](image)

For analytic purposes, filtering this network for the desired closeness above a certain threshold (e.g. > .7 for strong associations) might be a good thing to do. Other filtering options include filtering for specific topics represented by constituting terms and their
closest first order or second order term relations. The analysis might also focus on specific groups of documents. See Figure 3.4 for an example of a filtered network.

Furthermore, the size of the remaining network might still conceal its structure. Especially as typical latent semantic networks can easily span thousands of medium-frequent terms, such a rich network cannot be visualised easily.

![Figure 3.4 Filtered network.](image)

The closeness relations, however, can be clustered using, e.g., divisive clustering with Diana (Datta & Datta, 2006). This closeness hierarchy allows the identification of an appropriate cut-off level for a cluster analysis (appropriate for the level of analysis). Figure 3.5 shows a cluster hierarchy.

The result is the identification of component structures of the graph, which can again be visualised using conceptograms. Calculated from the well-known example used by Landauer et al. (1998) to present latent semantic analysis to the public, the corresponding latent semantic network is depicted in Figure 3.6. This figure clearly shows how the data cluster into two distinct groups.
3.4 The application: CONSPECT

3.4.1 Description
CONSPECT is the service prototype of WP4’s task 4.2 to put meaningful interaction analysis (MIA) of latent semantic networks into practice. The aim of this service prototype is to develop a tool that facilitates monitoring and inspection of conceptual development.
The processing pipeline from the brain to the screen is depicted in Figure 3.7. As competence is a potential for action, professional competence cannot be measured directly. Conclusions about the underlying competences can be drawn only from performance. Therefore, textual evidence that (ideally) explicates what a learner knows is collected in a first step in this processing pipeline. This input is then transcribed into latent semantics: the texts are folded into the dimensions of a latent semantic space and a representation of the learner text in the ‘language’ of the latent semantic space is created. Again ideally, this representation captures now more closely the meaning of the text (hence latent semantics).

This resulting representation is then presented to the user in the form of a visualisation or in the form of other types of user interfaces (for example a list of the core concepts touched on in the evidence material). Again ideally, this re-representation of the meaning expressed in the texts regarding their latent semantics allows the learner (or facilitator) to gain more effectively and more efficiently insight into what he knows – compared to having to work with the textual evidence alone.

By interacting with the interface, a user can discover new and previously hidden aspects in the latent semantic representations. One form of interaction is, for example, to re-arrange the layout of the conceptograms to discover whether certain clusters are in fact connected (or the visualisation merely suggests so as space is limited). Typical visualisation interactions help to scan the underlying complex data (the graph representing the latent semantics of the learner’s textual evidence).

Inspecting the conceptual, latent semantic representation with the help of this user interface mediated re-representation helps to reflect on the conceptual coverage and conceptual gaps. This supports the learner in making decisions about which area to focus on next or about which area needs more evidence (assuming that it is just a lack of evidence, not a lack of competence).
Figure 3.8 Meaningful interaction analysis deployed in CONSPECT.

Figure 3.8 summarises the steps involved in CONSPECT from the earlier parts of Section 3. Figure 3.9 is the SUM diagram, which shows the current implementation status of Task 4.2. Note that everything in the figure is implemented except for filtering and tagging.

The Maintenance module allows for managing spaces; it consists of three services: corpus management, LSA space configuration, and LSA space computation. The Access module has been implemented using openID. The Monitor module handles the functions that are most visible to the user. It contains several sub-modules. The Managing evidence sub-module handles the input. With it, the user can see what RSS feeds (learning blogs in Version 1) are available, can add feeds and can update feeds. The Managing representations sub-module prepares and displays a graph, i.e., a conceptogram. This sub-module lists the available conceptograms from which the user can select and allows the user to add and update a conceptogram. In addition, it contains the module that actually computes the information that will be displayed as a conceptogram. The Inspecting sub-module enables the user to look at a conceptogram. First, it identifies the graph the user wishes to see. Then it renders the graph using a lay-out algorithm and displays the conceptogram. A future version will allow for filtering complex graphs as shown in Figure 3.4. The Comparing sub-module identifies which two conceptograms the user wants to compare. It then calculates the agreement between them and renders the new, combined conceptogram. The Sharing module allows users to show their conceptogram to others, either one to one or to the public in general. It also offers a user the choice of accepting another user’s conceptogram. Version 2 will allow a user to tag a conceptogram with relevant concepts.
3.4.2 Required input

The current implementation uses PubMed abstracts as the training corpus. It also uses input from learners in the form of RSS feeds of relevant student blogs or learning diaries. In the next phase, we plan to use the human-marked forum conversations obtained by the University of Manchester.

3.4.3 Limitations of current implementation

A WP2 meeting held in December revealed several usability issues relating to the interface. The most important issue is that there is no legend that explains how to interpret a conceptogram. These deficiencies will be addressed in Version 2. However, a more substantive issue remains.

As currently implemented, the conceptograms show a model of language based on the background corpus. Therefore, the distances between concepts indicate the semantic similarity of words in the Real world Model. The conceptograms do not show distances in terms of the learner's model of the language. That is, there is no facility to indicate how a learner perceives the closeness of concepts and how those perceptions change over time. The conceptograms show how frequently the learner mentions a concept (by the size of the node) and which concepts the learner mentions or fails to mention with respect to another learner or source text (as in the intended learning outcomes example) by different
coloured nodes. But the conceptograms do not show variations over time in how the learner understands related concepts compared with each other.

We expect to add this functionality by, for example, displaying output based on comparing documents instead of terms. We would take two documents written at different times. We then produce two conceptograms, one for each document. One conceptogram would compare the learner's use of concepts in one document with the real world use. The other conceptogram would make the comparison using the other document. By comparing these two conceptograms, the user could see the difference over time.

**3.5 Usage example: learning about cancer**

In this section, we offer a coherent example of how CONSPECT can be used from the perspective of a learner. This illustrates with a sophisticated, yet simple enough user story, how learners could make use of the system in more reflective parts of their learning cycles. For simplicity, we have picked one particular topic to focus on: oncology – the study of cancer.

Evidence for creating a representation of what conceptual knowledge a learner may possess can be gathered from, for example, a learning diary or a forum discussion. This example uses real student blog postings by medical students.

The raw text of a learning diary is processed, using MIA, into a graph-like representation that reflects the student’s conceptual knowledge in a geometrical, latent semantic interaction model: the student's text is folded into a latent semantic space and within that space all words exert closeness or distance to each other according to their relatedness in meaning. The student’s text uses a specific vocabulary and thus stimulates only certain concepts in this latent semantic network. The resulting representation, a network-like structure, is a conceptogram.

**3.5.1 Interpreting Conceptograms**

CONSPECT, in its current implementation, offers two types of conceptograms - representations of a single source and combination conceptograms, where two sources are compared. Screenshots 3.1 and 3.2 show single source conceptograms; Screenshots 3.3 and 3.4 show combined conceptograms. These two types of conceptograms are interpreted differently.
A single source conceptogram shows concepts that appear in one text. The labels on the circles (known as nodes) are concepts written about in the source text. The size of the node indicates the relative frequency that a concept was mentioned in the text. The nodes are connected together by lines, also known as edges. The length of the edge between two nodes is a visual representation of how semantically similar are the two concepts represented by the nodes. (This semantic similarity is calculated from the LSA semantic space and is based on the background, or training, corpus.) The concepts are separated into component networks by clustering; each cluster is depicted with a separate colour. Each component shows concepts that are related semantically. The overlapping nodes are not significant; they are a result of the lay-out algorithm’s attempt to depict a great deal of information in a limited space.

Screenshot 3.1 displays a representation of a blog post from one student that includes an article about ‘pap smears’. As can be seen in the interactive diagram, there is a clustered
group about ‘cancer’ containing also concepts such as ‘lung’, ‘chemotherapy’, ‘breast’, or ‘prostate’. Additional groups can be identified: the student also blogs about ‘diets’ and ‘sex’ (in fact there was a longer article about sexually transmitted diseases).

Screenshot 3.2 Reference model constructed from textual description of intended learning outcomes.

Screenshot 3.2 is another example of a single source conceptogram. In this example, the texts describing the intended learning outcomes from an oncology course at an English university served as input. There is a purple cluster about cancer and concepts such as ‘prostate’, ‘breast’, or ‘chemotherapy’. However, the outcome descriptions stress that the course is also about patient care (orange cluster), pharmacology (green cluster, hidden), and research (blue cluster).

Screenshot 3.2 is an overview of the expected learning outcomes of an oncology course; Screenshot 3.1 is a snapshot of which concepts a particular student has written about.
Perhaps more informative than these types of conceptograms is a conceptogram that combines two others.

Screenshot 3.3 Agreement conceptogram comparing two student conceptograms.

Screenshot 3.3 compares the conceptograms of two learners. The gray nodes show the concepts covered by both learners. The red nodes show which concepts one student blogged about that were missing from the other student’s blog. The green nodes show the reverse. The sizes of the nodes and the distances between them have the same meaning as in the previous conceptograms.

When using LTfLL’s CONSPECT to compare this student against another, differences become salient: both students blog about cancer, one of them with an additional focus on ‘lung tumours’. Only one of them, however, deals with fractures caused by falling, whereas the other blogs about diets and other concepts. There are additional differences which can be explored by manipulating the diagram using drag & drop mouse operations.
Screenshot 3.4 Comparing a student conceptogram with the conceptogram for the intended learning outcomes of the oncology course of the University of Newcastle.

Screenshot 3.4 is another example of combining two conceptograms - the student blog and the intended learning outcomes. This combined conceptogram makes it clear that the student is providing evidence only about the hard core facts on cancer – thereby completely neglecting the care aspect. Similarly, attitudes, keeping up to date, and pharmacology are not mentioned.

4 How WP4 services relate to other services

4.1 Integration within WP2

In technical terms, the relation between implementations of WP4 services lies in the joint use of LSA and the WP2 provided LSA infrastructure. They meet especially over the
back-end for corpus management that allows administrators and tutors to create or sample specific purpose corpora and to manage the calculation of tuned latent-semantic spaces.

The current services implementation is in-line with the consortium’s decision by using REST styled web-services with defined interfaces. Therefore, it is easy for any service to call another service and use its output (XML) as a new input for further processing. All services follow the client/server principles and are modelled according to a three-tier architecture allowing for optimal subdivisions. We do not go into much more detail, because the technical architecture is extensively described in deliverables D2.1, D2.2 and D4.1b and we implemented the requirements completely.

Just as a brief summary of the WP2 integration, the services are designed and sub-divided into three layers: data layer (database, file system), service layer (the web-services), and a presentation layer (web pages, widgets). For the data layer we use a MySQL database to store texts and the file system to store uploaded learning materials (as PDFs, DOCs etc.). Web-services are programmed in R and PHP and are called using REST requests. Data is given back using a custom XML syntax, which is described in a service directory document. For the presentation layer, we render the output as a normal HTML website. AJAX calls (using the jQuery library) invoke the web-services, fetching the data, and generating the output. We have developed our own HTML framework in PHP where we embed the widget-styled outputs (accessible through http://augur.wu.ac.at/v1/wp4.1). Therefore, we are able to test the widget-styled user interfaces and can also easily widgetise and integrate them into a learning environment. We are currently working on the integration of the widgets by using Wookie as a widget engine and Elgg as the learning environment and have done some preliminary tests to ensure that the services can be integrated as expected.

Relation to services of other WPs (WP4.2 and WP5.2) is technically also done by providing an interface for NLP tasks (stemming, stop word filtering, tokenizing etc.) and especially for LSA space managements and calculations. For version 2 of the services, it has to be investigated if this interface can be used by WP5.1 as well. The learning materials management service is also a good candidate for a shared service as WP6 would profit from it and - the other way round - WP4 can use annotated and personally tagged resources to improve positioning and the conceptual development tasks.

4.2 Integration within WP3 (threads)

As explained in the CAI report, WPs 4, 5 and 6 fit the part of the Stahl cycle diagram that describes interaction between the cycles of personal and collaborative knowledge building. The lower left corner of the diagram shows the cycle of personal understanding, which can start with a tacit pre-understanding influenced by personal knowing. The right part of the diagram depicts how the social process of interaction with people and with our shared culture influences the individual’s understanding. Although in the diagram personal understanding and social knowledge building are separated, it is only a matter of representation, they can only be separated artificially.

Task 6.1 is based on the same assumption that the stable knowledge of a CoP is encoded within an ontology. There the ontology is used to index learning materials and to
facilitate the retrieval of such materials. One obvious step after the positioning of the learner on the basis of task 4.1 is the usage of the services of task 6.1 to locate appropriate learning materials for further reading. In addition, task 6.1 service can benefit from the qualitative analysis (task 4.1) output that can we used as an intermediate step to discover lexicalisations for the ontology concepts.

In D4.1 we explained that task 4.1 knowledge rich approach and task 6.2 can be integrated by means of using folksonomies representing a democratic consensual knowledge instead of ontologies that are built mainly by experts. In addition the overall positioning service can be integrated to the WP6 service for specifying learner position and find suitable materials online.

Version 2 of the positioning service implementing discussion forums will be capable of interoperating with the 5.1 services as both services’ functionalities overlap in relation to the analysis of chats and conversation to identify conceptual learner coverage. Task 4.1 service can therefore enrich 5.1 services by incorporating the linguistic usage dimension into the analysis.

In task 4.2, we find possible integration points for bringing together the formal language work within LTfLL, especially with WP6: ontology creation, annotation, and presentation could go hand in hand with the analytics provided to inspect the conceptual development of learners.

With task 5.2, the work of both WP4 tasks already shares a common basis in the joint use of LSA. All LTfLL tasks collect some sort of evidence of learning in digital form and task 4.2 could principally serve as a common data sink and feedback logbook with feedback type delivery for all. Future work on integration will show.

5. Verification and evaluation plans

5.1 For language technologies (in task 4.1)

While validation of pedagogical relevance and usability of 4.1 services will be reported in LTfLL D7.3 report (D7.3) this section discusses the technical specifications for the ongoing verification of the effectiveness of the languages technologies implemented in task 4.1 services.

The positioning service evaluates learner texts by classifying evidence of proper language usage and conceptual coverage as sufficient, unclear or insufficient. Recall and precision are the most used evaluation measures in classification. Those measures evaluate a system judgement against human judgments (gold standards). Manning & Schütze (1999) define those measures as follows in terms of selected and non-selected documents (items) that are relevant targets for the system as defined by human experts (gold standards). 

Precision measures the proportion of selected items that the system got right and recall measures the proportion of the target items that the system selected. Additional measures
resulting from the combination of recall and precision are also widely used e.g. $F$-measure.

As we have already explained in D4.1, verification is done by comparing training results obtained with a traditional bag of words LSA configuration and with an alternative configuration that uses bag of phrases. Cross validation is then used over the chosen configuration.

We are verifying task 4.1 approaches on the level of concept annotation within learner texts. First, we select representative texts according to their quality as indicated by marks given by tutors. These texts are annotated manually by two annotators and also automatically by the service. In this way, we measure the inter-annotator agreement using statistical measures for categorical data e.g. kappa coefficient (Cohen, 1960) and we calculate the precision and recall of the automatic annotation. By having texts with a different quality, we are also able to determine for which kind of texts each of these approaches is most reliable.

**5.2 For CONSPECT (task 4.2)**

We will undertake ongoing evaluation of the conceptual development service, during Alpha and Beta development, in three areas - technical specifications, or accuracy, pedagogical relevance, and usability. Validation activities for version 1 of CONSPECT will be reported in detail in D7.3. Verification activities for the service are summarised, below, as a series of experiments.

Any tool to help learners is useful only if it is accurate. There are (at least) two aspects of accuracy that must be evaluated. The first area is the semantic similarity of concepts calculated from the latent semantic space. The second area relates to concepts extracted from a text. The next sections describe several experiments to test these two areas of accuracy.

*Experiment 1* determines whether humans group concepts in the same way as does CONSPECT. It is a type of card-sorting evaluation. We generate lists of descriptive concepts of a text using CONSPECT. We ask our participants to arrange the descriptors into groups so that each group contains strongly associated descriptors. Through these groups, we could see if the connections between the concepts generated by CONSPECT agree with those connections given by the participants. During development, we may have ideas for improving CONSPECT which would then create a new LSA space. Even if the descriptors change with the new space, we still should be able to reuse the human grouping.

*Experiment 2* determines if humans agree with the descriptors that CONSPECT gives to a text. We generate a list of concepts given by CONSPECT to a specific text. We then give the text to our test participants to read and ask them to rank each LSA-descriptor on a scale from 1 to 5 (5 = very descriptive, 1 = absolutely non-descriptive). We add a couple
of distracters to evaluate the judgements against chance (should be significantly better rated than chance).

Finally, Experiment 3 is a Priming Experiment; it is the most problematic of the three experiments because it depends on us obtaining appropriate software to carry out. Assuming we can locate such software, we would use a latent semantic space to get a rich set of terms in a particular domain along with their LSA-similarity (=association strength). Then we conduct an associative priming experiment to see, how strongly they are related to our participants. Associative priming uses one keyword (e.g., ‘table’) to stimulate the participant. It then notes the reaction time to complete an abbreviated keyword, say ‘cha’. As table and chair are more strongly associated, the probability that the participant would extend the abbreviated keyword to ‘chair’ (and not to, say, ‘character’) is higher than in a control group without the prime. There are ways to use reaction time to evaluate the association strength. The purpose of the priming experiment is to verify to what extent the distance between concepts measured by LSA agrees with human participants.

Even if the associations from the LSA space turn out not to be good enough, the results of the priming experiment can be reused later. Even if the target area is not perfectly defined by the original LSA space, the terms selected should be good enough to sketch it, and the association strengths from the priming experiment stay stable. The validity has to show that convergence remains among the terms that are strongly associated by the participants and divergence exists where they are not.

Of course, the validity is influenced by carefully selecting the target group to be representative of the users of the tool.

6. Conclusions and future work

Task 4.1 positioning service implements two analyses of the learner texts. An analysis of language usage (quantitative and qualitative) based on a knowledge poor approach and an analysis of learner self assessment based on a knowledge rich approach. While the first analysis reflects learner integration into the relevant CoP, the second reflects the learner’s coverage of relevant concepts. The language usage qualitative analysis calculates the distance between bags of distinctive phrases used with learner texts and instruction materials. The phrases are extracted from high quality texts (e.g. learning materials, learner texts, etc.) by means of analysing (qualitatively) the linguistic usage of the relevant CoP when referring to specific areas of expertise and then identifying those phrases within learner text under evaluation. The analysis to be implemented for version 2 of the knowledge rich approach identifies the learner’s sentiment towards his own knowledge of specific required concepts and calculates the percentage of those that are covered. The service also provides support for building the reference corpus by means of a non-parametric statistical test used to evaluate instructional material as prototypes for language use and conceptual coverage.
In this report we have described the implementation of a text management service that facilitates the upload of texts into the service and the integration of the positioning service with other service from LTfLL project (5.1 and 6.2). In the next cycle of the project, task 4.1 will focus on the validation of Version 1 of the positioning service and the implementation of Version 2 where a semi-automatic knowledge rich approach will be deployed.

In relation to the conceptual development services implemented by task 4.2, the language technologies are already set up and the first prototype service for monitoring conceptual development is delivered. The next and immediate phase relates to the validation of the service. Progress with this respect will be reported along with the deliverables of the validation WP7. Data gathered should ideally enable us to conduct further pseudo experiments re-using this as a human benchmark to evaluate the effectiveness of the service.

Several extensions for the task 4.2 service version 2 are already planned. Most notably these are two extensions: linking with literature and enhancing interaction methods. Functionality to link concepts with literature and to allow for browsing through these relations is foreseen. Most promising seems to be the approach to use the relation to the PubMed abstracts in the background corpus for establishing this relation, as this rich set of 9 million documents is a promising repository of learning objects.

We have plans to allow for enhanced interaction between the user interface representation and the underlying complex data: one aspect thereof is to link with sources to explain where concepts originate. Another aspect is to provide multiple levels and granularities of aggregation: the mantra overview first, zoom and filter, details on demand will be the guiding principle for these extensions.
References


Appendix 1. User interface

The user interface for our first version is intended to provide the basic means for our learner-positioning approaches, providing automated feedback to the user regarding his answers to a questionnaire for a course or community of practice, as well as supporting the tutor in his grading task. This not only requires an interface to our positioning services, but also a basic course management system. The design is structured with a graphical header above the currently selected content and a footer below (Screenshot A1.1). However, in the following screenshots only the content sections (or relevant fractions of it) will be shown.

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Description</th>
<th>Domain</th>
<th>Language</th>
<th>Progress*</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Medical Course</td>
<td>UNIMAN short tests</td>
<td>English</td>
<td>0/6</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>IT Basics</td>
<td>IT Basisausbildung</td>
<td>German</td>
<td>0/10</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>IT Basics (Eng)</td>
<td>IT Basic Knowledge</td>
<td>English</td>
<td>5/10</td>
<td></td>
</tr>
</tbody>
</table>

*Shows the number of solved/total questions by the user in this course.

Screenshot A1.1: Student course management.

For the first version we have implemented two roles, the tutor and the student. When first browsing to the page, the user gets redirected to a login screen. Screenshot A1.1 displays the initial view when a student is logged in. There, all the available courses including relevant information are displayed for the user. The interesting part here is the progress column, where the share of already solved or answered questions for the course is displayed. In the next step, the student selects a course by clicking the “view” button. This view is split into several subsections. First, the general course information is displayed at the top. The next section is the “Learning Material” subsection. This lists the available learning material uploaded by the tutor. The student may download these documents to prepare for the course questions.
Screenshot A1.2: Student questions list.

Screenshot A1.2 shows the questions for this course, another subsection of the course view. The green mark shows whether the student has already solved the question (in this case received a “grade” of 100%) or not. To view and answer one of these questions the student has to click the corresponding “view” button.

Screenshot A1.3: Student answer view.

As illustrated in Screenshot A1.3, the selected question is then displayed in detail followed by the student’s answer. It is possible that the student has already created an answer before, and may now be editing or resubmitting the answer. The grade or question
coverage can be awarded by the tutor and can then be seen by the student in this view, as well. From this point, the student may alter his answer text and click “submit” to resubmit the text and await the tutor’s response - or request automated live feedback from the system before that. Which opens the “Live Feedback” frame as displayed in Screenshot A1.4.

Screenshot A1.4: Live Feedback view.
In this case the user gets 3 different kinds of feedback: First, a quantitative score of 85% is awarded to the answer by a fully automated LSA-based algorithm. At this stage we calculate the semantic proximity (cosine) of the learner text to already graded answer texts and compute a cosine-weighted average of their grades. Secondly the detected or and missing concepts, as processed by the “Knowledge Rich” approach, are displayed. Finally the qualitative feedback in the form of a list of “Distinctive Phrases” helps the user identify phrases from his text that were mainly used in either high or low graded documents. Even though the concepts found in the “Knowledge Rich” feedback and the automatically detected “Distinct Phrases” are produced independently, the results are mostly quite consistent. For the first version, the “Knowledge Rich” approach is only available for one course due to data integration issues that have to be solved.

With regard to integrating the different services in the project for the next version, the answers could also for example be displayed in a concept map as developed by work package 4.2. In that case, the answers concept map could also be matched to an expert’s concept map derived from the existing gold standard answers, which have been graded with the maximal score.

Tutors have a similar interface with a few adjustments. For example in addition they may also add, edit and delete courses, learning material and questions. They are the ones responsible for finally grading the student answers, while having the same live feedback tools at their disposal. In contrast to the student, the tutor interface allows the tutor to list all student answers to a specific question and helps him to quickly find ungraded or resubmitted answers. Tutors also have a simplified overview of the status of the semantic space and are able to retrain the semantic space on the fly. No deep knowledge of the algorithms behind the system (for example LSA) is required for the tutor in this view.

However, in order to achieve better results with our LSA based feedback algorithms, we have implemented a system where one can calculate semantic spaces for every course individually in only a few relatively simple steps. This option can be used by a language expert to make sure each of these spaces contain the necessary contextual information to provide valid judgments about the semantic proximity of documents in that specific field. The implementation maintains the goal of maximal flexibility by providing a vast array of preprocessing options as well as an open database back end for the text data and corpus management. This is linked to an interface that is able to work with sparse matrices, allowing for very efficient and low memory consumption and thus for the calculation of larger and more complete semantic spaces.

The interface we have built for the semantic space management is structured similarly to the course management. Each space has a control section listing the parameters that space has been built with (or will be built with the next time a build is invoked). These are mostly preprocessing parameters ranging from classic natural language processing options such as stemming, stop-word filtering and lower case normalization to fully customizable tokenisation functions (word tokenisation by default), various and combinable weighting options (such as IDF weighting), minimum collection and
document term frequencies and length filters, as well as the code to select the corpus for the semantic space. The code parts (tokenizing and document sourcing) have to be written in R, the programming language used to calculate the semantic spaces. To enable a faster comparison of documents the already graded documents may be folded into the semantic space beforehand. By default, whenever a new course is created, a semantic space with default parameters is prepared and may be recalculated and trained whenever new course data is available.

This interface is in a development stage and is undergoing constant improvement. In addition, it will grow/be adjusted as demand for additional functionality rises. For now, it provides a rather convenient way of calculating latent semantic spaces and making them available for our LSA based web-services. Naturally, this service can be adopted seamlessly for other tasks using LSA algorithms such as 4.2 or 5.2, as well.

Another area of work here (investigated in version 2) will be adding a system that lets the user alter or create custom feedback algorithms and that provides a means to automatically evaluate these algorithms in terms of common retrieval measures such as precision or calculating the average deviation of an algorithm for machine scores versus the scores awarded by tutors/humans. This not only will provide for example valuable data on which scoring algorithm works best, but also which semantic space with what parameters works best.

From a technical point of view the HTML user interface is mainly generated by querying our XML webservice via AJAX calls, loosely coupling a set of websites together in one coherent interface. This way further development/enhancements are very flexible as well as a possible widget integration is ready from the start.
Appendix 2. Text management system

The web-based text management system consists of two databases. The course management (Figure A2.1) and the training materials databases. The first database allows tutors to create courses, store learner answers and provide feedback to learners by grading answers and writing comments. For each course, tutors can upload various learning materials (documents), define a number of questions, link those questions to learner answers and relevant learning material. Once a sufficient number of graded answers is available the system can automatically grade the student’s answer, providing automated feedback (qualitative in the form of topic-distinct phrases and quantitative in the form of a grade based on statistical calculations of term frequencies and LSA) to the student or provide feedback for the tutor, who provides the final grade for the answers.

Figure A2.1: Course management database design.

The second database stores additional text materials used for training the services. For example, we have added tables containing 9 Million Medline documents that may be filtered according to their MeSH (Medical Subject Headings) headings to build the base corpora for medical-related questions. Other examples are a table holding ECTEL papers that may be used for information technology or e-learning related corpora, a table
comprising the discussions on a medical online forum as well as a collection of French paragraphs taken from the “Le Monde” newspaper.

These tables have not been streamlined to a common format since they require very specific metadata. For example the Medline documents tables include MeSH information with which the documents have been categorised by experts, which helps with proper filtering. Also the medical discussion table holds communication flow information concerning which text was a response to which other text, which texts are the root documents, etc. The negative side of having “custom” tables like these is that they can hardly be integrated in an automated environment on the fly, which may be needed in a production environment. At this stage of development however we prefer the more flexible approach of generating these data repository and corpora tables in a custom way, specifically tailored to our ever-changing data requirements for our experiments.

This basic structure may be used for other services as well, such as building concept maps from student answers (task 4.2) or providing learning material and course domains for summary writing (task 5.2). The latent semantic spaces generated by the system may be used by any LSA-based workpackage. The highly flexible interface for the space generation provides easy access to the document collections we have stored in our database and lets the user generate tailor-made latent semantic spaces for his NLP task.