Project Deliverable Report

D4.1 – Positioning Design

<table>
<thead>
<tr>
<th>Work Package</th>
<th>WP4 Positioning the Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>4.1, 4.2</td>
</tr>
<tr>
<td>Date of delivery</td>
<td>Contractual: 01-11-2008</td>
</tr>
<tr>
<td></td>
<td>Actual: 22-12-2008</td>
</tr>
<tr>
<td>Code name</td>
<td>D4.1</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Type of deliverable</td>
<td>Report</td>
</tr>
<tr>
<td>Security (distribution level)</td>
<td>Public</td>
</tr>
<tr>
<td>Contributors</td>
<td>Gaston Burek, Adriana J. Berlanga, Marco Kalz, Isobel Braidman, Alisdair Smithies, Fridolin Wild, Petya Osenova, Kiril Simov, Bernhard Hoisl, Lothar Lemnitzer, Slavi Stoyanov, Peter van Rosmalen, Jan Hensgens, Maria Regan, Jan Van Bruggen, Gillian Armitt</td>
</tr>
<tr>
<td>Authors (Partner)</td>
<td>UTU, OUNL, UNI-MAN, WUW, IPP-BAS, AURUS-KTS</td>
</tr>
<tr>
<td>Contact Person</td>
<td>Gaston Burek</td>
</tr>
<tr>
<td>WP/Task responsible</td>
<td>Gaston Burek</td>
</tr>
<tr>
<td>EC Project Officer</td>
<td>Mr. M. Májek</td>
</tr>
<tr>
<td>Abstract (for dissemination)</td>
<td>This report explains the issues of positioning learners in a knowledge domain, to recommend learning materials or courses to follow, and diagnosing learners’ conceptual development, to provide formative feedback and recommend remedial actions. It also discusses how Language Technologies can be used to perform these two tasks in a (semi) automatic way. To this end, the report presents an outline of the existing tools, methods and resources for the analysis of ePortfolios (task 4.1), and an overview of the process of acquiring and analysing data for measuring conceptual development (task 4.2). Furthermore, it presents a description of an initial set of experiments and test results, a plan for extending existing tools for ePortfolio analysis, as well as a description of the pilot scenarios that will be tested in the next cycle of the project.</td>
</tr>
<tr>
<td>Keywords List</td>
<td>Learner Positioning, Diagnosing Conceptual Development, ePortfolio, Latent Semantic Analysis, Knowledge Rich Approaches, Language Technologies, LTfLL</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>We would like to thank Philippe Dessus for providing valuable comments on previous versions of this deliverable.</td>
</tr>
</tbody>
</table>

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

1. Introduction ........................................................................................................................................... 4
2. Positioning the Learner .......................................................................................................................... 5
   2.1 ePortfolios ........................................................................................................................................... 9
      2.1.1 ePortfolios to Support Learner Development ........................................................................ 10
      2.1.2 ePortfolios and the Medical Profession ................................................................................. 11
      2.1.3 The Role of ePortfolios in Undergraduate Medical Education ........................................... 11
      2.1.4 The Role of ePortfolios in Computer Science Education ......................................................... 13
      2.1.5 Positioning the ePortfolio Learner ............................................................................................. 13
   2.2 Language Technologies for Positioning the Learner ........................................................................ 14
      2.2.1 Latent Semantic Analysis ........................................................................................................... 14
      2.2.2 LSA Based Knowledge Modelling for Positioning ................................................................. 17
      2.2.3 Measuring Textual Coherence .................................................................................................. 17
      2.2.4 Data Sets for Evaluating (P)LSA Based Positioning ............................................................... 18
      2.2.5 Knowledge Rich Approaches: State of the Art .................................................................... 20
   2.5 Extending Available Infrastructure .................................................................................................... 23
      2.5.1 Building an Infrastructure for (P)LSA-Based ePortfolio Analysis and Positioning ............ 24
      2.5.2 Extending Knowledge Rich Approaches Tools ....................................................................... 25
   2.6 Future Work ......................................................................................................................................... 29
3. Diagnosing Learner’s Conceptual Development .................................................................................... 29
   3.1 Theoretical Background ................................................................................................................... 32
      3.1.1 Learners as Experts, Intermediate or Novices (in Medicine) .................................................. 32
      3.1.2 Diagnosing Conceptual Development: The Case of Medical Schools .................................... 34
      3.1.3 The Process of Diagnosing Conceptual Development .............................................................. 36
   3.2 Description of the Work to Be Done ................................................................................................. 40
      3.2.1 Existing Methods and Tools Approach ...................................................................................... 40
      3.2.2 Latent Semantic Approach ........................................................................................................ 44
   3.3 Future Work ......................................................................................................................................... 47
4. Conclusions ........................................................................................................................................... 48
5. References ............................................................................................................................................ 49
6. Appendixes ........................................................................................................................................... 54
   6.1 Appendix – 1: Description of the Pre-pilots ....................................................................................... 54
      6.1.1 Task 4.1. LSA Based Learner Positioning in Relation to his ePortfolio and Learning Materials ............................................................................................................. 54
      6.1.2 Task 4.2. Monitoring Learner’s Conceptual Development in the Medical Domain ................. 58
   6.2 Appendix – 2: First Experimental Results Task 4.1 ......................................................................... 61
   6.3 Appendix – 3: First Experimental Results Task 4.2 ......................................................................... 65
   6.4 Appendix – 4: Test of the ALA-reader, CMAP and KNOT Tools .................................................. 66
   6.5 Appendix – 5: Review of Systems Performing Sentiment Analysis ........................................... 72
   6.6 Appendix – 6: Text Available for Analysis ....................................................................................... 74
Executive Summary

LTfLL work package 4 focuses on two independent but connected issues:

1. Determining the learner's position with regard to courses in a study domain, or in a set of possible courses, to provide the learner with the 'best' suitable material to achieve their learning goals.

2. Determining the conceptual development of a learner related to a particular expertise area, to provide them with formative feedback and remedial actions to reduce knowledge gaps.

Traditionally, standard assessment methods are used to address these issues. For lifelong learners, however, these methods do not apply. These learners have a heterogeneous background, in terms of competences and knowledge, and might follow a combination of different educational providers to achieve their learning goals. In these situations, positioning and diagnosing learner’s conceptual development will require more resources, in order to provide learners with adequate and accurate support, and prevent tutors’ workload from getting out of hand.

In this report we claim that these issues can be reduced if the learner’s position and diagnosis of their conceptual development are performed as (semi) automatic tasks, applying the latest advances in the research domain of Natural Language Processing, and particularly Latent Semantic Analysis, as well as in the application of ePortfolios for lifelong learning. To this end, this report attempts to answer the following questions:

- What is the scope of positioning and diagnosing conceptual development problems?
- Which theoretical background is relevant to develop a solution to tackle these problems?
- How will this background be considered to develop solutions for these problems?
- What are the solutions that will be developed?
- What technologies and methods will be used or developed to implement the solutions?

To answer these questions, this report is organized in two main sections: positioning the learner (task 4.1) and diagnosing learner’s conceptual development (task 4.2). The former discusses the importance of ePortfolios in Education and, particularly, for positioning the learner, as well as the Language Technology approaches (i.e. Latent Semantic Analysis and knowledge rich based) to be implemented for ePortfolio analysis. It also presents first experimental results as well as the work that has been done to extend existing technologies. The latter section presents the theoretical background, the rationale behind, and the processes of diagnosing learner’s conceptual development. This section also depicts the two approaches that will be conducted to evaluate to what extent Language Technologies perform better than or can extend other methods and tools to measure conceptual development; showing, moreover, some initial results and the work planned to
establish the first prototype. Finally, the last section of the report draws conclusions. The document also includes several appendixes that detail different issues covered in this report, such as pre-pilot scenarios, first experiments, systems, and text available for the pilots.
1. Introduction

Learners enrolled in educational institutions, traditionally, share characteristics such as a common learning goal, level of knowledge, and background. Conversely, lifelong learners are self-directed learners who have different learning goals and heterogeneous profiles formed by different (formal and informal) educational backgrounds. These learners need services that support them in finding appropriate paths (to be followed in order to achieve their learning goals), in a way that they know their position with respect to the courses and programs multiple educational providers offer.

Educational providers, that want to widen their educational profile and offer services for lifelong learners, will face some challenges to support lifelong learners’ education. Firstly, the assessment of each learner, in terms of her knowledge and competences (i.e., learner’s position with respect to the course prerequisites), will require more time if they are assessed as lifelong learners instead of “traditionally”: standard assessment methods no longer apply due to the lack of homogeneity on the learners’ background. Secondly, tutors will be confronted with an increase of workload, as they will need to assess the position of a larger number of learners and to decide, almost on an individual basis, which resources fit better each learner’s needs. Thirdly, while learning, learners will need to receive specific formative feedback that shows what they know, what conceptual gaps they have, and which remedial actions they should take to overcome these gaps.

In order to mitigate these difficulties, the work package 4 (WP4) approach is to develop learner support services that (semi) automatically position the learner with regard to the learning goals and learning materials and, based on that, recommend the optimal (or a suitable) learning path, and give the learner formative feedback about their study progress and the best suitable material to help them to achieve their learning goals (e.g. successfully complete a course, just undertaking a particular concept, etc.).

Positioning the learner requires mapping the characteristics from a learner’s learning history and knowledge gaps to the characteristics of learning materials and curricula. Learner history is typically contained in the learner’s ePortfolio. This work package, therefore, builds on and advances the state of the art to more specific areas of ePortfolio, gap analysis, and monitoring and diagnosing conceptual development. The state of the art in this area is still characterised by tedious data collection, evaluation and assessment, especially when informal learning is concerned (Merrifield et al., 2000). Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) promises to automate the analysis and interpretation of the data. LSA creates a mathematical model in which textual descriptions of both domain knowledge and learner’s knowledge can be projected. It is still an open question whether LSA can cope with these issues in a multi-lingual context and can help to provide semantically rich interpretations and feedback.

WP4 is divided into two tasks. In task 4.1, positioning the learner, techniques to match ePortfolio input and learning objects (courses) with the goal of imposing an order on the learning materials and to offer an optimal path for each individual learner are investigated. To this end, a linguistically enriched variety of LSA, in which predicate argument structures instead of bags of words are used as input, will be used. This approach, which has been applied successfully to the problem (Burek et al., 2007), is
similar to the one explained in (Dessus, 2004) in the context of Intelligent Tutoring Systems. Therefore, the project will also draw from state of the art in language technologies and develop it further. A particularly promising line of research from the field of applied natural language processing is Recognition of Textual Entailment (Dagan et al., 2006), where NLP is used to assess the logical relation between a text and a hypothesis (i.e., does the text entail the hypothesis or not?).

In task 4.2, conceptual development, techniques to diagnosing and monitoring learners’ conceptual development are investigated. The proposed solution is based on the observation (Boshuizen and Schmidt, 1992; Boshuizen and Schmidt, 2004) that a learner’s conceptual development is closely reflected in the textual utterances learners express as part of their evolving domain knowledge. More precisely, the concepts used and their relations expressed by novices and experts change through time in a systematic, experience based fashion. Based on this observation, existing methods and tools to diagnose conceptual development will be evaluated and compared with a “pure” LSA, in which only a direct LSA technique is used.

The rest of this document is structured as follows: section 2 describes the work carried out in the context of task 4.1, positioning the learner. It starts describing the issue of positioning the (lifelong) learner in a domain: they need recommendations of the best suitable material of courses to follow in order to reach their learning goals, and depicts the purposed approach to solve this issue. After, the section discusses the theoretical background of the use of ePortfolios in education, the language technology approaches (i.e. LSA and knowledge rich based) to be implemented for ePortfolio analysis and positioning, and presents first exploratory experimental results. Section 3 describes the work carried out in the context of task 4.2. It starts describing the issue of monitoring and diagnosing conceptual development of (lifelong) learners to provide formative feedback, and depicts the purposed solution for tackling this issue. After, the section presents the theoretical background and the rationale behind and the process involved. The section also depicts the two approaches that will be conducted to evaluate to what extent Language Technologies perform better than or can extend other methods and tools to measure conceptual development; showing, moreover, some initial results and presenting the work planned to establish the first prototype. Finally, the last section draws conclusions on our findings.

2. Positioning the Learner

When (lifelong) learners want to achieve learning goals or develop further their competences, they will consider the different options educational providers offer to enrol in a curriculum or select a course that better fits their needs.

These learners, however, come from a variety of backgrounds, might have different learning goals and different prior knowledge; the traditional model of ‘one-curriculum fits all’ will no longer apply. Learner’s prior knowledge, for instance, can be of a formal nature (certified exams, certificates) in which case standard admission / exemption procedures may apply. In other cases, as in non-formal learning, such standard procedures are not available. These differences imply that some learners will benefit from following certain courses or lessons within a curriculum, while others definitely will not.
In order to provide learners with the 'best' suitable courses and materials to be followed, their individual “position” in a curriculum, or in a set of possible courses, should be calculated. In this context, the course or material is considered “suitable” whenever it provides learners with the knowledge that is necessary for achieving learner’s learning goals.

To address this issue the learner’s prior acquired knowledge should be considered. Accreditation or Recognition of Prior Learning (APL/RPL) provides a procedural solution for identifying the prior knowledge of learners in formal and informal education (Merriefield et al., 2000). In this case, evidence of prior knowledge is provided by the learner that needs to be compared to public or domain knowledge (see Table 1). In this way, a plan of a learning path, from learner’s background to the desired outcomes, can be generated. In this learning path, sections of the curricula may be exempted. Only after this intake, the learner can be assigned to particular courses.

<table>
<thead>
<tr>
<th>Learner</th>
<th>Objectives indicated by learner</th>
<th>Evidence indicated by learner</th>
<th>Materials indicated by assessors</th>
<th>System Recommendations</th>
<th>Assessors decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Competence 1</td>
<td>Self assessment</td>
<td>Course documents</td>
<td>No automatic</td>
<td>Accept yes/no</td>
</tr>
<tr>
<td></td>
<td>Competence 2</td>
<td>Certificates</td>
<td>Course documents</td>
<td>No automatic</td>
<td>Accept yes/no</td>
</tr>
<tr>
<td>John Johnson</td>
<td>Competence 3</td>
<td>EPortfolio documents</td>
<td>Materials or documents accepted/rejected</td>
<td>Automatic</td>
<td>Accept yes/no</td>
</tr>
<tr>
<td></td>
<td>Competence 4</td>
<td>Formal assessment</td>
<td>Assessment samples accepted/rejected</td>
<td>Automatic</td>
<td>Accept yes/no</td>
</tr>
</tbody>
</table>

Table 1. Competences

In this report we present the design of a service that is based on the automatic analysis of learner learning history and provides an automatic RPL given a particular curriculum. This curriculum may be offered by formal (e.g. University) or non-formal (e.g. a cooking club) education providers.

The initial description of such a service is as follows: firstly the service recommends a list of personalised curricula in addition to curricula offered by a single education provider. The recommendation is based on the analysis of a learner’s written material during their learning history or alternatively it is based on the learner learning goal as explicitly specified by the learner in writing. From this list, the learner chooses a curriculum (personalised or not personalised). Secondly the service tells the learner what evidence (mainly documents produced by the learner) are the strongest (although maybe
D 4.1 Positioning Design

not sufficient) to demonstrate prior knowledge on the section of the chosen curriculum. Thirdly, the service recommends assessors (who have the responsibility of assessing learners’ prior knowledge) on whether learners have enough prior knowledge on a specific subject for him/her or skip a particular section of the curriculum. Fourthly, the service produces feedback on the recommendation provided. When an assessor evaluates the service recommendation for each section of the chosen curriculum by accepting or rejecting it, he/she writes a text supporting each of his/her decisions. The service will save the supporting text written by the assessor. It is assumed that eventually, the service will learn given the previous decisions of the assessor to produce automatic feedback on the evidence presented by the learner.

Next, the service will save the learners’ presented evidence with the corresponding label according to the assessor decision (i.e., accepted evidence or rejected evidence). Finally, the service generates an optimal sequence of learning materials that individual learners need to study as indicated by an advisor. That sequence of learning materials is optimal in the sense that it considers the individual learner’s prior knowledge, learner’s learning goals, the available curricula, and the coherence among learning materials. The service will sequence the learner materials by means of calculating LSA based measures of similarity among the texts (e.g. sentences, paragraphs, etc.). The resulted sequence consists in texts ordered according their similarity. The first text materials in the sequence must correspond to learning material that best match the prior knowledge of the learner.

The service will provide automatic recommendations about competences where learner-produced materials are relevant (e.g. competence 3 and 4 as defined in Table 2). The service will compare evidence of prior knowledge (from learner’s ePortfolio) against previously presented evidence together with learning materials playing the role of gold standards. Initially, assessors will make their own independent decision on whether the learner needs to study or to skip a learning object. The quality of the automatic recommendation provided by the service will depend on the number of cases previously presented to the service. We assume that eventually the service will become an expert and will guide the assessor decisions with more reliability. To provide the automatic recommendations the service needs to perform the following intermediate actions as indicated in the DoW:

- Model the public or domain knowledge using various Language Technology techniques
- Model the private knowledge of the student in a similar way (ePortfolio analysis)
- Compare student knowledge to domain knowledge; compute gap (APL)
- Find learning materials to cover the gap
- Sequence material
- Advise material for study
- Provide automatic feedback

1 NB. Competence 1 is outside the scope of work package 4 tasks, the DoW mentions that the data to be used in task 1 are documents included in learners’ portfolio (see Table 2).
D 4.1 Positioning Design

<table>
<thead>
<tr>
<th>Learner</th>
<th>Objectives indicated by learner</th>
<th>Evidence indicated by learner</th>
<th>Materials indicated by assessors</th>
<th>Automatic Recommendations for assessors</th>
<th>Assessors(^2) Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Johnson</td>
<td>Topic 1</td>
<td>Documents 3, 5 and 8</td>
<td>Materials or documents accepted/ rejected</td>
<td>Advice sufficient =&gt; skip, Advice unclear =&gt; study, Advice insufficient =&gt; study</td>
<td>Accept yes/no</td>
</tr>
<tr>
<td>Topic 3</td>
<td>Documents 1, 4 and 7</td>
<td></td>
<td>Materials or documents accepted/ rejected</td>
<td>Advice sufficient =&gt; skip, Advice unclear =&gt; study, Advice insufficient =&gt; study</td>
<td>Accept yes/no</td>
</tr>
<tr>
<td>Topic 3</td>
<td>Documents 2, 3 and 4</td>
<td></td>
<td>Materials or documents accepted/ rejected</td>
<td>Advice sufficient =&gt; skip, Advice unclear =&gt; study, Advice insufficient =&gt; study</td>
<td>Accept yes/no</td>
</tr>
</tbody>
</table>

Table 2. Types of evidence for automatic recommendations in competences 3 and 4 (as defined in Table 1)

As described in the DoW, the combination of knowledge rich approaches and LSA (PLSA) is considered as a solution to automatically analyse the learner’s presented evidence and to calculate their position in the curricula. The private knowledge of the learner, demonstrated in the evidence, needs to be evaluated by comparing it against the public knowledge, represented by the learning materials.

Kalz et al. (2007) discuss different cases in which these approaches are relevant. In the first case LSA is relevant because only text is available. The authors mention that in such a case the learner wants to enrol in a curriculum that has not been annotated with standard formal descriptions of competences. In addition learner’s evidence on their prior knowledge consists in raw text. In the second case, the learner has a standardised ePortfolio with formal descriptions of competences and the curriculum is formally described according to the same standards i.e. with formal description of competences. EPortfolio standard specifications (e.g., IMS LIP, IMS e-Portfolio) can be used as standards for the formal descriptions. In the third case, a learner ePortfolio is formally described with standard descriptions belonging to a competence ontology and the curriculum is also formally described with a standard belonging to the same competence ontology.

Alternatively to using probabilistic and traditional LSA to provide automatic recommendations over evidence and curricula belonging to the first case (as depicted in Fout! Verwijzingsbron niet gevonden.), the service we are envisioned will semi-automatically annotate the documents and curricula as specified in the third case. Once

\(^2\) In case of non-formal learning the role of assessor can be fulfilled by a peer or by the persons studying themselves.
the documents and curricula are annotated, the system will use ontology labels knowledge rich based approaches to generate equivalent recommendations.

As part of the theoretical background needed to develop this solution, the next section explores the use of ePortfolios to support learner’s development in general and in particular for the medical and computer science education, and discusses their significance to solve the position problem described before.

2.1 ePortfolios

Cotterill (2007) states that, “In general, an ePortfolio is a purposeful collection of information and digital artefacts that demonstrates development or evidences learning outcomes, skills or competencies. The process of producing an ePortfolio (writing, typing, recording etc.) usually requires the synthesis of ideas, reflection on achievements, self-awareness and forward planning; with the potential for educational, developmental or other benefits. Specific types of ePortfolios can be defined in part by their purpose (such as presentation, application, reflection, assessment and personal development planning), pedagogic design, level of structure (intrinsic or extrinsic), duration (episodic or life-long) and other factors. EPortfolio applications allow the owner to share specific parts or views of their ePortfolio online and support feedback and dialogue....”

The Centre for Recording Achievement describes ePortfolios in similar terms, identifying the following attributes as a typical set of services (centre for Recording Achievement):

- A means of accessing personal information held in distributed databases;
- A means of selecting sets of items for a specific purpose and making connections and associations between them and with specified standards/rubrics;
- A means of allowing other specific individuals to view any given selection, and controlling the time within which they are allowed to view;
- Guidance to support review and choice, reflection and action planning
- A means of sharing and collaborating with individuals and communities.

The theory underpinning the selection and use of ePortfolios in educational and professional contexts, like ePortfolios in general, draws upon a broad range of learning theories including, but not limited to, reflective, experiential and constructivist approaches to learning. The development of ePortfolios, however, has developed through the complex requirements of lifelong learning, often in a work place environment. Mason et al. (2004) point out that selection of items for ePortfolios is facilitated by ready organisation of material. Re-ordering and revision of items is also made easier through the electronic medium, which also allows sharing of selected items with groups of learners. Driessen et al. (2007) identify in their comparison of web-based portfolios (e.g., ePortfolios) to paper-based portfolios that ePortfolios enhance students’ motivation, were

---

more user-friendly for mentors, and delivered the same content quality compared with paper-based portfolios.

While in Mason et al. (2004) authors describe at length the affordances of, and requirements for, extensible personal and community services, portability and adaptability of content, Barrett (2003) describes collection, selection, reflection, projection, presentation as associated with ePortfolio use in literature in terms of supporting educational process.

From a formal assessment perspective, Chang (2001) describes ePortfolios as providing the student with “authentic, reflective, interactive and individual features,” all of these attributes having advantages over examinations and computer-assisted, multiple choice forms of assessment. Mason et al. (2004) conclude, “As a method of assessment, the ePortfolio builds independence and learning-to-learn skills, which are necessary components for the lifelong learner.” (pp.727).

2.1.1 ePortfolios to Support Learner Development

Portfolios are widely incorporated into professional development for many professions; nursing, teaching, psychotherapy and counselling, law, engineering are all examples of careers which require mandatory (e)portfolios. The content of such portfolios are defined by the appropriate regulatory bodies for each profession and are required for scrutiny and evidence during regular professional appraisals. Its role has therefore developed from its earlier function as an indicator of an individual’s learning and achievements to one in which the evidence that it contains clearly demonstrates the knowledge and competences required to progress within professions.

The Centre for Recording Achievement identify that ePortfolios can provide learners with the following services to support learners critically reviewing their progress in relation to action planning for professional development4:

- A means of importing a set of criteria/standards/outcomes (as with assessment)
- A means of creating and storing a plan (and ideally, a means of sharing that plan with others: certainly with the tutor/mentor/coach, maybe with peers)
- A means of accessing that plan at a later point in time
- A means of recording experiences and achievements as they happen
- A means of accessing ongoing records of experiences/achievements and selecting those which may be relevant to the criteria/standards/outcomes being reviewed
- A means of reminding the user (mentee, learner, appraisee) of timelines and due dates.

EPortfolios, in the context of lifelong learning, have been identified as having huge potential to support constructive profiling of learners’ achievements and experiences. Learners, through regular recording, review and reflection; maintain an evidence base to

---

demonstrate their individual development and engagement with learning experiences. ePortfolios, thus, develop confidence and promote ownership of responsibility for learning by providing not only a means to establish career aspirations but also the means to identify and record the personal journey undertaken.

2.1.2 ePortfolios and the Medical Profession

Many of the key features of portfolio are exemplified by the use of them in medicine, both in the undergraduate course and for qualified professionals. The introduction of (e)portfolios into medicine is relatively recent, but they are now required by many countries as essential evidence for the development of professional standards by individual clinical practitioners, as exemplified in the USA, Canada and the UK. For medical practitioners in the UK, the portfolio must provide evidence that the individual is competent, abides by high ethical standards, maintains the standards the public and the medical profession expect and is thus fit to practice medicine, as stipulated by the General Medical Council (GMC), the profession’s national regulatory body. This more stringent approach to portfolios and to personal and professional development was necessitated as a response to occurrences of poor medical practice and of medical catastrophes, which diminished public trust in the medical profession, and is encapsulated by the GMC’s current Good Medical Practice guidelines. This has several implications for the education of medical students. First, many practicing doctors have no experience of maintaining a portfolio. They are unclear about how to do this and about which evidence is appropriate. Introduction of personal and professional development portfolios to the undergraduate curriculum accustoms students to maintaining portfolios and this has implications for their learning, before they qualify as clinicians. Second, there is awareness that engagement of medical students with the issues of good medical practice, included in the portfolio, may be an indicator of fitness to practice as a qualified clinician. The manner in which students maintain their portfolios may therefore provide some of the evidence on which decisions to allow students into the medical professional could be based. Thirdly, newly qualified doctors, in their first two years of postgraduate training, termed Foundation Years, are required to maintain a standard portfolio, which demonstrates their skills and competencies as well as their ability to reflect critically on their experiences. The portfolio is an essential document for the assessing Foundation Year training and medical students are expected to be familiar with this process.

2.1.3 The Role of ePortfolios in Undergraduate Medical Education

In the UK, all university students are expected to maintain a record of their personal and professional development, but also for medical students, the GMC stipulates that personal and professional development ePortfolios are an essential component of the undergraduate curriculum. The medical students’ ePortfolio is therefore related to the GMC guidelines.

The notion of a record of development, which not only contains evidence of achievements and ability to perform skills competently, but also demonstrates a student’s

---

6 For more information see [http://www.foundationprogramme.nhs.uk/pages/home/key-documents](http://www.foundationprogramme.nhs.uk/pages/home/key-documents)
progress as an independent reflective learner, is a complex concept both for students and for educators to deliver successfully. Care must be taken to ensure that students appreciate its relevance to their learning experience, so clarity and a clear basic structure are important prerequisites for the undergraduate ePortfolio. In the UK the ePortfolio originally introduced at the University of Dundee Medical School has been used as a model, but the seven components of the GMC’s Good Medical Practice guidelines are also gaining in popularity. Whichever basic structure is used a ePortfolio is expected to contain:

- Evidence of engagement in activities that promote professional development (i.e. that is signed by the appropriate authorities, is dated and is verifiable)
- Evidence that the student is acquiring key clinical skills and the level of competency reached in those skills
- Critical self analysis of the student’s own strengths and limitations and how they are building on the former and addressing the latter
- Evidence that the student is responding to the feedback given to him/her by those assessing development in skills and clinical practice
- Demonstration of how the individual student is developing their own learning strategies and clear indications that they can formulate their own learning plans, especially at key stages in the curriculum
- Evidence that they are developing an understanding of disease processes, which they apply to their clinical learning situations
- Clear indications that students understand and develop competencies in key features of modern medical practice e.g. teamworking, communications, governance of medical practice and evidence based medicine
- Critical thinking and reflective learning, which underpins many of the above features of the ePortfolio.

A key feature of the success of ePortfolios in the undergraduate curriculum is that they are actively supported by faculty either through an individual mentoring system or through small group interactions supported by tutor facilitators. This support enables students to discuss progress and obtain feedback on their development.

An important issue which has stimulated considerable debate concerns the assessment of ePortfolios. There is evidence that this can be achieved by inspection and grading solely of the written ePortfolio. Other experience with this approach indicates that there may be significant variation between assessors, although this may be overcome by clear training guidelines. An alternative approach is to interview the student with their ePortfolio and this has the advantage that progress can be discussed and the student regards the ePortfolio as something indicative of their own stage of development rather than a series of pieces that are graded. Usually in this approach, ePortfolios are either satisfactory or unsatisfactory, so there is less opportunity for assessor variation.
2.1.4 The Role of ePortfolios in Computer Science Education

One objective of the project is to apply the techniques and services, which we will develop, to more than one domain and language. We decided to choose computer science as the second domain for the following reasons: a) as one outcome of the LT4EL project (www.let.uu.nl/lt4el/), we have access to a large number of learning objects from the domain of computer science and in various languages; in addition we have access to lexica which have been derived from these learning objects and to an ontology connecting all the terms through language-independent concepts; b) some of the partners have experience in teaching computer science.

In contrast, the situation with ePortfolios for this domain is far from ideal. In contrast to e.g. medicine, the use of ePortfolios is not an established practice in science and engineering (cf. Bhattacharya et al., 2006). Within the consortium there are no partners actively using ePortfolios in Computer Science as part of their teaching. We will therefore endeavour to acquire ePortfolio matching our domain from outside. This has a drawback that the construction of the ePortfolio has to be supported and guided. However, it also gives a clear advantage, since in this way we can explore, design and negotiate ePortfolio input that has not yet been fixed but is open for requirements derived from the techniques proposed.

In any case we are convinced that the practice of using ePortfolios will become more widespread, and therefore we see an increasing number of users of our service, which uses ePortfolios in order to position the learners and suggest individual learning paths to them.

2.1.5 Positioning the ePortfolio Learner

Overall, the preceding discussion and explanation of ePortfolios emphasises the importance of the evidence they contain for the development of the individual learner or practitioner. This is exemplified in the undergraduate medical ePortfolio where evidence is available from relatively restricted sources at the beginning of the course, indicative of the learner’s initial experiences. By the end of the curriculum, however, the ePortfolio covers a comprehensive range of material, demonstrating the learner’s progress in terms of cognitive development, critical thinking, professional identity and acquisition of skills and competencies. In the final year, a typical ePortfolio will contain evidence of research projects, Student Selected Components (see Appendix 6 for details), attendance at learning events and activities, learning plans agreed with educational supervisors, acquisition and understanding of essential skills, as observed by senior expert clinicians, discussions of clinical case management with senior clinicians and reflection on key incidents and experiences of specific importance to the individual. Most of these will be accompanied by feedback from the senior clinician concerned. An important feature of the ePortfolio at this stage is the response of learner to the feedback that they have received. Thus, a personal advisor or tutor will be able to advise the individual learner by taking an overall view of the evidence, identifying, with the learner, their overall strengths and weaknesses and enabling them to formulate a clear plan for the next stage of their learning trajectory.
In summary, the “ideal” ePortfolio for the medical student is one which genuinely helps and supports the student’s own personal and professional development. It demonstrates his/her development as a learner, both in terms of understanding and acquisition of skills and competencies, based on ever improving critical and reflective thinking and learning, thereby preparing the student for their life as a medical practitioner. In the context of the proposed automatic analysis (i.e. knowledge poor and knowledge rich based approaches), technology has the capability to inform and guide students through composite tasks of the “ideal” ePortfolio.

2.2 Language Technologies for Positioning the Learner

As specified in the LTfLL package 4 description of work Latent Semantic Analysis (LSA), and its probabilistic variant PLSA, together with knowledge rich approaches will be used for the automatic analysis of ePortfolios aimed to positioning the learner. The rest of this section presents the theoretical background of these technologies.

2.4.1 Latent Semantic Analysis

LSA has widely used in applications for automatic assessment of essays, the provision of feedback and selection of suitable materials according the learner degree of expertise in specific domains. For instance, the Intelligent Essay Assessor (IEA) assesses automatically essays using a semantic space built from materials on the topic to be evaluated. In (Foltz et al., 1999) the authors report that the IEA rating performance is close to the one of human raters.

LSA is an algorithm applied to approximate the meaning of texts, thereby exposing semantic structure to computation. LSA combines the classical vector space model, well known in computational linguistics, with singular value decomposition (SVD), a two-mode factor analysis. Thus, bag-of-words representations of texts can be mapped into a modified vector space that is assumed to reflect semantic structure.

LSA is intended to enable the analysis of the semantic structure of texts. The basic idea behind LSA is that the collocation of terms of a given document-term vector space reflects a higher-order, latent semantic, structure, which is obscured by word usage (e.g., by synonyms or ambiguities). By using conceptual indexes that are derived statistically via truncated singular value decomposition, this variability problem is believed to be overcome (Deerwester et al., 1990).

In a typical LSA process, first a document to term matrix $M$ is constructed from a given text base of $n$ documents containing $m$ terms. This text matrix $M$ of the size $m \times n$ is then resolved by the singular value decomposition into the term-vector matrix $T$ (constituting the left singular vectors), the document-vector matrix $D$ (constituting the right singular vectors) being both orthonormal and the diagonal matrix $S$.

These matrices are then reduced to a particular number of dimensions $k$, giving the truncated matrices $T_k$, $S_k$ and $D_k$ : the latent semantic space. Multiplying the truncated matrices $T_k$, $S_k$ and $D_k$ results in a new matrix $M_k$ which is the least-squares best fit approximation of $M$ with $k$ singular values. $M_k$ is of the same format as $M$, i.e., rows represent the same terms, columns the same documents.
To keep additional documents from influencing a previously calculated semantic space or to simply re-use the structure contained in an already existing factor distribution, new documents can be folded-in after the singular value decomposition. For this purpose, the add-on documents can be added to the pre-existing latent semantic space by mapping them into the existing factor structure. Moreover, folding-in is computationally a lot less costly, as no singular value decomposition is needed.

To fold-in, a pseudo-document vector \( m \) needs to be calculated in three steps (Berry et al., 1995): after constructing a document vector \( v \) from the additional documents containing the term frequencies in the exact order constituted by the input text matrix \( M \), \( v \) can be mapped into the latent semantic space by applying the following equations:

\[
\begin{align*}
d &= v^T T_k S_k^{-1} \\
m &= T_k S_k d
\end{align*}
\]

Thereby, \( T_k \) and \( S_k \) are the truncated matrices from the previously calculated latent semantic space. The resulting vector \( d \) represents an additional column of \( D_k \). The resulting pseudo-document vector \( m \) is identical to an additional column in the text matrix representation of the latent semantic space.

When applying LSA, a process is executed that typically involves several (optional) steps and involves various data-types created as an output of these steps. Below we provide a list of entities and processes involved in a typical LSA process.

- **Term**: The word as it is written in a document.
- **Corpus**: The collection of documents containing texts that consist of terms separated by punctuation marks.
- **Text matrix**: A representation of the document collection in matrix format: the cells contain the frequency, how often a particular term appears in a specific document. Terms are the rows, documents the columns. By transforming a corpus to this representation format, documents are treated as so-called bag of words, where the term order is neglected.
- **Latent-Semantic Space**: When applying a singular-value decomposition (SVD) to a text matrix, the matrix is resolved into the term-vector matrix \( T \) (constituting the left singular vectors), the document-vector matrix \( D \) (constituting the right singular vectors) being both orthonormal and the diagonal matrix \( S \) (Berry et al., 1995). These partial matrices are then truncated in order to reflect strong associations, eliminate noise, etc. The set of these three, truncated partial matrices \( T_k, S_k \) and \( D_k \) is called 'latent-semantic space'. A latent-semantics space can be converted back to the text matrix format.
- **Folding in**: To keep additional documents from changing the structure of a latent semantic space, documents can be folded into the previously calculated space. Thereby, \( T_k \) and \( S_k \) of the space are re-used and combined with a text matrix constructed over the new documents. See Wild and Stahl (2006) for more details.
- **Dimension**: When truncating the partial matrices from the SVD, a particular number of the highest singular values are retained. This is called the dimensionality of the latent semantic space.

- **Distance / Similarity**: Within a text matrix, various methods can be applied to measure the distance (or – the other way round – similarity) between terms, documents, or terms and documents. One method is, e.g., to use the measure the cosine of the angle between two column-vectors in order to calculate the similarity between two documents. A high cosine value is equal to a small angle between the vectors, thus indicating high similarity.

- **Vocabulary**: All terms used within a corpus form the vocabulary of this corpus. The vocabulary has a certain order to ensure that additional text matrices can be constructed that can be appended to an existing text matrix.

### 2.2.1.1 Probabilistic Latent Semantic Analysis (PLSA)

PLSA (Hofmann, 1999) introduces a statistical methods to the classic linear algebra based LSA. While LSA uses a linear algebra approach for reducing the dimensionality of textmatrices by mean of the singular value decomposition (SVD), PLSA reduces the textmatrix dimensionality by means of mixture decomposition based on a the aspect model.

The aspect model takes advantage of the notion that each document represented is a mixture of concepts. The model incorporates the probability of document $d$ from $D=\{d_1, d_2, d_3, ..., d_N\}$ be representative of a specific concept $z$ belonging to $Z=\{z_1, z_2, z_3, ..., z_k\}$ and the probability that concept $z$ is expressed by term $w$ from to $W=\{w_1, w_2, w_3, ..., w_M\}$.

The latent model relates probabilistically an unobserved class variable $z$ belonging to each data sample. Joint probability model of documents and words $P(d, w)$ where $d$ and $w$ are independently conditioned on the value of $z$.

Relationship between LSA and PLSA is given by the following equation:

$$P_{PLSA}(d, w) = \sum_{z \in Z} P(d|z) P(z) P(w|z)$$

where the reduced LSA decomposed matrices $T_k$, $S_k$ and $D_k$ with rank $k$ can be defined as follows in terms of PLSA,

$$T_k = (P(d_m|z_k))_{m,k}$$
$$S_k = \text{diag}(P(z_k))_k$$
$$D_k = (P(w_n|z_k))_{n,k}$$

where the joint probability model can be written as a product $P = T_kS_kD_k^T$ being $T_k$ and $D_k$ independent of each other.

The parameters for a $k$-concepts PLSA model are $k$ multinomial distributions of size $n$ and $m$ mixtures over the $k$ hidden concepts. This gives $kn + km$ parameters and therefore linear growth in $M$. 

---

**DATE:** 2008-212578

**PAGE:** 16
For fitting the model, tempered expectation maximization (TEM) can be used as alternative to the Maximum Expectation (EM) algorithm to avoid PLSA over fitting the data. The results reported on the performance of PLSA when compared to LSA showed a significance improvement in performance.

2.4.2 LSA Based Knowledge Modelling for Positioning

As explained in Van Bruggen et al. (2004), LSA-based positioning requires creating a latent-semantic space from text documents that model learners’ and public knowledge on a specific subject. Texts documents from learner’s ePortfolio (e.g. written material from the learner’s own production, descriptions of learning activities completed by the learner, etc.) can be used to model individual learners’ knowledge. Educational materials (e.g. textbooks, articles, etc.) can be used to model public knowledge.

For LSA purposes the texts need to be rich in the use of the specific subject related terminology. LSA relies on the use of rich terminology to characterise the semantics of natural language. To this end, corpora and data sets are essential. In Van der Vegte et al. (2007) authors explain that three different types of data sets are needed for LSA. A large collection of text that defines the underlying language and a smaller collection defining the domain language are needed for building the text matrix. In addition, the authors mention the test corpus as the third type of data set. This data set consists of a set of documents with an associated set of similarity relationship (e.g. a grading scale) between them. Here a fourth type that corresponds to the training set needs to be added. In order to improve accuracy with the testing set, it is necessary to train a set of model parameters beforehand (similarity measure, number of LSA dimensions, similarity value to use as threshold, weighting scheme). This procedure is called cross validation. To train the parameters it is necessary to evaluate different configurations for those parameters using a training set and choose the parameters configuration that gives the best results. This configuration will be then evaluated with the testing set that consists in unseen samples during the training and provides a notion about the system capabilities to generalise.

In relation to the size of the text collections, Quesada (2007) mentions that “a corpus is big enough when, for any learning experience added, the probability of adding a new type is so low that it is negligible”. The conclusion, therefore, is that the parameters can only be optimal for a specific class of types and, as a consequence, the data set to be used should be representative of that specified class of types.

2.4.3 Measuring Textual Coherence

In addition to measuring similarity between whole documents (e.g. learner’s essays and learning materials chapters), a positioning service requires the evaluation of sections or sentences that may belong to the same document or to different documents. More specifically to evaluate the coherence in learners’ written documents the service is required to measure the semantic similarity between sentences of the same documents.

Wiemaer–Hasting and Graesser (2000) use a measure of coherence to understand how one sentence relates to other and then provide feedback on that basis. Coherence indicates that contiguous texts are more similar than distant texts. As reported by Van Bruggen et al. (2004) coherence LSA-based measures were found to be stable thorough pages within
psychology books. This observation suggests that learning materials recommended to learners can be sequenced on the basis of the degree of coherence among them.

According to Foltz et al. (1998) in order to comprehend a text readers create a connected representation of textual descriptions of concepts. The readers related those concepts by means linking. The process of linking concepts involves determining and maintaining coherence between texts segments. Among others causal relationships and coreference are two of elements of discourse that contribute to textual coherence.

As traditional (P)LSA treat texts as bag of words, current (P)LSA based approaches for evaluating coherence are not capable of modelling the semantics of more than one concept per text (sentence, paragraphs, document, etc). That is a significant limitation of the current approaches given that two or more concepts can be described within a single text e.g. a sentence can describe different concepts and how they relate to each other by means of a subject, verb, and object.

In addition, those approaches are also limited in that they are not capable of recognising directionality in causal relationships. In (Burek et al., 2007) the authors present a triple based LSA approach for quantifying the strength in the textual entailment relationship between texts. This approach makes use of semantic similarity based alignment of n-grams represented as vectors in a semantic space. By allowing the analysis for fine grained semantics this approach calculates a set of similarity measures between the semantically related constituents of the sentence structure (i.e. subject, verb and object).

### 2.4.4 Data Sets for Evaluating (P)LSA Based Positioning

For evaluating the performance of the positioning service we plan to build a data set consisting of texts pairs. The first text describes a specific topic and will be generated by domain experts. Experts may generate more than one description per topic.

The second text in the pair represents a sample of evidence of learner’s knowledge on the topic specified by the first text. As samples containing such evidence are expensive and scarce, we plan to use surrogate samples instead. We will harvest the surrogate samples from domain specific online resources using (P)LSA based models of learner’s prior knowledge.

Among the materials incorporated into available ePortfolios, which are suitable for modelling learner evidence of prior knowledge on specific topics, are text outputs from discussion. In these, students discuss issues related to the safe prescribing of medicines, based on a series of clearly defined learning objectives. Preliminary analysis of the discussion threads identified six main topic areas:

- The role of medical students in patient safety
- The role of the ward pharmacist in preventing serious medication errors
- The critical points in medicines management where serious errors can occur
- The Swiss cheese model of accident causation
- The minimum core knowledge, skills and attitudes required to prescribe safely
The legal consequences of negligent or reckless prescribing / administration of drugs.

Individual students' contributions have been selectively sampled based on a set of criteria that will provide a good quantity of content. Each contribution is presented in a format that includes the student identifier, the topic area that is covered and a grade (i.e. Excellent, Good, Fair, Poor) indicating the effectiveness with which the student has covered the domain topic. For each of these, three model answers or golden standards will also be provided, for comparison.

2.4.4.1 Recall and Precision
Poisoning can be seen as a classification task where evidence is classified as sufficient, unclear or insufficient (see Fout! Verwijzingsbron niet gevonden). Recall and precision are the most used evaluation measures in classification. Those measures evaluate a system judgment against human judgments (gold standards). Manning and Schütze (1999) define those measures as follows in terms of selected and no selected documents (items) that are relevant targets (see Figure 1) for the system as defined by human experts in the gold standards.

**Precision** measures the proportion of selected items that the system got right.

\[ \text{Precision} = \frac{tp}{tp+fp} \]

**Recall** measures the proportion of the target items that the system selected.

\[ \text{Recall} = \frac{tp}{tp+fn} \]

![Figure 1. Relationship between true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn), adapted from Manning and Schütze (1999)](image)

The **F-measure** (van Rijsbergen, 1979), a widely used measure, results from the combination of recall and precision.

\[ \text{F-measure} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Recall} + \text{Precision})} \]

The **F-measure** is a weighted average of the precision and recall, where the highest possible value it can reach equals 1 and worst value equals 0.
2.4.4.2 Exploratory Results
We have carried out a set of LSA experiments as part of task 4.1 show case validation (for more details see Appendix 1). In those experiments we compared machine judgments against human judgments. LSA was used to calculate the machine judgments. The objective of these experiments was to test if LSA can learn by training the classifiers. We cross validated the classifiers using different learning rules. The initial experimental results, using small size training and testing set, seems to indicate that LSA can indeed train its classifiers (see Appendix 2). We will run more experiments using larger data sets to validate these preliminary results.

2.4.5 Knowledge Rich Approaches: State of the Art
As alternative to the (P)LSA approach knowledge rich approaches are considered here in two ways: (1) a separate module for learner positioning; and (2) a module which provides information for the LSA-based methods to do their job on conceptual level. In our work on the positioning of the learner with the help of the knowledge rich approaches we will rely on the ideas reported in (Kalz et al., 2007). They discuss the notion of learning network. According to it, learner’s competence can be automatically compared to a set of concept evidences of the target competence. Our goal is to achieve an ontology-based positioning where the learner competence is represented by a learner’s competence ontology and a curriculum competence ontology. However, reliable competence ontologies are still missing. Thus, in our work we will rely on domain ontologies which reflect the knowledge part of the learner’s competence. The ontological analyses of the learner’s ePortfolio and the textual description of the relevant curriculum will be an approximation of learner’s competence and curriculum competence. In order to introduce some first steps in the evaluation of the learner’s knowledge degree we will evaluate the usage of this knowledge represented within the ePortfolio. This evaluation will be done via the techniques of the sentiment analysis. After analysing the curriculum description and the ePortfolio, we will use several approaches to compare the extracted conceptual information. In the rest of this section we discuss the envisaged analyses in the process of ePortfolio analysis.

Knowledge rich methods rely on analysis of the text by using knowledge sources outside of the text, such as lexicons, ontologies, grammars. In the case of ePortfolio analysis, the result from it is used as an evidence of the learner competence and knowledge in the domain. Within the framework of work package 4 we consider several types of text analysis (also their applications to learning tasks, and more specifically the positioning of learner with respect to a curriculum description). These text analyses include: (1) ontology-based semantic annotation, (2) discourse segmentation, (3) lexical chains approach to disambiguation of concept annotation and (4) sentiment analysis for evaluation of the concept usage in the text. The combination of these analyses has to explicate the conceptual content of the ePortfolio which to be used for positioning of the learner. The ontology-based semantic annotation relates the text of a ePortfolio to the conceptual information in the ontology. The discourse segmentation facilitates the creation of lexical chains and the sentiment analysis. The lexical chains identification supports the disambiguation of the ambiguous terms and phrases within the text. The sentiment analysis determines the attitude of the learner to the concepts explicated within
the ePortfolio. At the end, the result of the whole set of analyses presents a classification of the concept usages in the text as known and unknown. This classification will be used for the comparison to a conceptual representation of a curriculum. Here are some recent references to relevant works.

- (Galley and McKeown, 2003) present the idea of the lexical chains, which trace the cohesion within the texts. The automatic establishment of lexical chains is outlined in (Mihalcea, 2007).

- (Wolf and Gibson, 2006) propose 11 coherence relations, which successfully segment the discourse within the texts. According to Schauer (2000), 15 to 20 percent of coherence relations are signalled by some kind of conjunction but not all of them are unambiguous.

- The various levels of sentiment analysis scope are described in (Moilanen and Pulman, 2007) and (Liu, 2008), among others. It is often underlined that adding knowledge rich features improves the results in sentiment analysis. For example – (Moilanen and Pulman, 2007), (Kennedy and Inkpen, 2006), (Kim and Hovy, 2006).

As mentioned above, knowledge rich approaches are usually connected with the availability and the usage of knowledge rich data bases (see Figure 3), such as ontologies and lexicons. The ontologies reflect the conceptualisations in some domain of interest. For example, DAML ontology library, SWOGGLE, or LT4eL ontology. These ontologies have to be connected to an upper ontology in order to cover in better way the general knowledge. For example, DOLCE, SUMO, SIMPLE. The most famous knowledge rich lexicons are the so-called wordnets (WordNet, EuroWordNet, BalkaNet, SIMPLE). Such resources are exploited for semantic annotation of documents and/or for semantic retrieval. For better semantic annotation and its usage in positioning of the learning task we consider discourse segmentation and sentiment analysis methods as relevant.

Within LT4eL project an ontology-to-text relation was defined (Simov and Osenova 2007; Simov and Osenova, 2008). We briefly present this relation here. We assume that the ontology is the repository of the lexical meaning of the language. Thus, we have started with a concept in the ontology and we searched for lexical items and non-lexical phrases that convey the content of the concept. There are two possible problems here: (1) there is no lexical item for some of the concepts in the ontology, and (2) there are lexical items in the language without a concept representing the meaning of the lexical item in the ontology. The first problem is overcome by allowing in the lexicon also non-lexical (fully compositional) phrases to be represented. The second problem is solved by extension of the ontology. The lexicon items are then mapped to grammars. We call them concept annotation grammars. These grammars relate the lexicon to the text. Such a mapping is necessary as much as lexical items and phrases from the lexicons allow for multiple realisations in the text. Thus, they require some additional linguistic knowledge in order to disambiguate between different meanings of some lexical item or phrase. Figure 3 depicts the elements of the model.
We have been using the relations between the different elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is relied on for the concept annotation of the text. In this way we established a connection between the ontology and the texts. The relation between the lexicon and the ontology is used for definition of user queries with respect to the appropriate segments within the documents.

Another direction of the knowledge rich methods is the discourse analysis. As a benchmark, the work of Wolf and Gibson (2006) can be considered. The authors present 10 coherence relations. The advantage of their work is that they succeed to collapse the big number of possible relations into a small set of operational relations.

Sentiment (opinion) analysis can be specified broadly as a kind of analysis that aims to determine automatically the attitude (sentiment, tone, polarity) of a speaker / writer with respect to a certain topic. Usually this kind of analysis is being opposed to the standard fact-based analysis and in the same time it is rendered as a classification task. It is commonplace to evaluate the sentiment of an opinion on a two value-scale: negative or positive. When free of subjectivity, the text is regarded as neutral.

Much work is done for detecting negative or positive judgments, but sentiment analysis is not only about the more general polarity of an opinion, it is also about identifying the opinion holder, the object (the topic) that have been evaluated, the type of propositional attitude expressed (belief – think, believe, assume, emotion – hate, adore, etc.), and the strength of the polarity. Of course, more features can be added.

The type of texts subject to this kind of information retrieval are numerous, just to give some examples: customer reviews on all kinds of products, brands, reviews on cultural events, opinion polls. Here are some freely downloadable corpora with different domain-specific texts: (1) MPQA Opinion Corpus, containing 535 news articles collected during the 2002 NRRC Workshop on Multi-Perspective Question Answering (Wiebe, 2002); (2)
polarity dataset, containing 1000 positive and 1000 negative movie reviews (Pang and Lee, 2004).

Sentiment analysis is concerned with two levels of granularity: sentence and document level, the second estimated as too coarse for most applications (Liu, 2008). According to Moilanen and Pulman (2007) though, while sentiment classifiers work well with a large input (e.g. a 750-word movie review), the results for sentential and subsentential units – clauses or noun phrases, are not satisfying. Taking into account linguistic features, such as valence shifters (for example negation) intensifiers, gradable adjectives, patterns, semantic role labelling and syntactic structure, adding a level of compositionality do improve the analysis in terms of accuracy (Moilanen and Pulman, 2007; Kennedy and Inkpen, 2006; Kim and Hovy, 2006, amongst others).

The prevailing majority of techniques use some form of machine learning, supervised and semi-supervised. The most common and basic approach to sentiment classification is keyword-based, starting from a list of sentiment indicators or clues prepared manually, semi-automatically (relying on WordNet of FrameNet) or acquired by machine-learning (Rimon, 2005). Other types of elements used in different algorithms are: Semantic Orientation (SO) – Pointwise Mutual Information (PMI), support vector machines, maximum entropy, naïve Bayes, latent semantic analysis (LSA) and so on (for a detailed overview see Liu, 2008). For a description of some systems performing sentiment analysis see Appendix 5.

In this knowledge rich approach for learner positioning we will rely on the reported works. We will integrate the above technologies in a common processing module in order to explicate the conceptual content of the ePortfolio and the curriculum description. The explication of the conceptual content will be done via annotation of the text part of the ePortfolio and the curriculum description. This annotation could be used as input for different tasks. Firstly, the concept annotation will be used to find the position of the learner with respect to the curriculum and to select appropriate learning materials to cover the gaps discovered by the method (see below for more details). Secondly, the concept annotation within the text could be used as an input for (P)LSA methods. Concepts could substitute the terms within the textmatrix.

To evaluate the performance of this knowledge rich approach for positioning we plan to use comparable data sets and validation method as the ones proposed for the (P)LSA based approach.

2.5 Extending Available Infrastructure

With the purpose of implementing the proposed service for ePortfolio analysis and positioning, we plan to extend available infrastructure for LSA and knowledge rich based analysis. Below we provide a description of our plan for extending the existing tools.
2.5.1 Building an Infrastructure for (P)LSA-Based ePortfolio Analysis and Positioning

Wild (2005) describes the LSA package for the statistical language and environment R. R7 is an environment for statistical computing. The software is compatible with UNIX platforms, Windows and MacOS.

The design of the (P)LSA based service for ePortfolio analysis and positioning builds up as an extension of existing LSA infrastructure, the R package (for more details see Wild, 2005). In a typical R package LSA process a textmatrix is constructed from the input corpus. The textmatrix can (but does not need to be) weighted using one of the various weighting schemes provided. Then, the singular-value decomposition is executed over the textmatrix and the resulting partial matrices are truncated and returned. The number of dimension to keep can be obtained using a set various recommender routines.

In case that additional documents are to be folded into the existing latent-semantic space, again a new text matrix is constructed using text matrix re-using the vocabulary from the first one. Again the resulting textmatrix can be weighted (eventually re-using the global weights of the first textmatrix). The resulting text matrix can be folded into the existing latent-semantic space, thereby re-using the truncated left-sided and the diagonal partial matrices of the SVD. In this case, the result is directly a text matrix.

Looking more closely at the R package text matrix routine, it can be seen that several text sanitizing and pre-processing steps are embedded in the text matrix generation routines: the routine included means to convert the terms to lower case, simple routines for stripping XML tags, automatic removal of punctuation marks and some other special characters, and trimming of multiple white spaces. Furthermore, stop words can be filtered (by providing stop word lists) or a controlled vocabulary can be deployed. Furthermore, frequency filters can be applied to delete terms below or above a certain frequency threshold (within a document or within the corpus) or outside a certain term length range. Terms consisting purely of numbers can be removed automatically. Also all terms can be reduced to their word stems by mean of using Porter's snowball stemmer (Porter, 1980). The package is open-source and available via CRAN, the Comprehensive R Archive Network.

We plan to extend the routines already implemented within the R-package. Those extensions consist in new routines for (P)LSA on large sparse matrices, for latent-semantic space, parameters estimation, etc., and routines for triple based LSA as described in (Burek et al., 2007).

The proposed (P)LSA based ePortfolio analysis and positioning service will articulate the functionalities provided by the extended R package in addition to other functionalities including:

- Storing input and output data and documents (e.g. training and testing data sets, learners' ePortfolio documents, learning materials, etc.)

7 http://www.r-project.org/
• Annotating and handling of learning materials and ePortfolio documents according to previous recommendations given by the service.
• Evaluating evidence by analyzing learner’s ePortfolio document and determine if there are sufficient evidence of learner’s prior knowledge on specific topic. This functionality evaluates (P)LSA ePortfolio documents and learning materials
• Skipping or studying recommendations of sections of a curriculum. Those recommendations are based on the analysis of ePortfolios documents, curricula and learner’s learning goals
• Sequencing of the learning materials recommended for study.

2.5.2 Extending Knowledge Rich Approaches Tools

As mentioned above, additionally to (P)LSA based functionalities, the service will provide knowledge rich ePortfolio positioning functionalities. For that purpose we plan to extend the CLaRK system (Simov et al., 2001) originally implemented with the aim of minimizing human work during the process of corpora creation. CLaRK will provide service’s functionalities for calling external programs when they are necessary for some specific task. CLaRK is implemented in Java and the necessary functional interface will be provided.

The knowledge rich methods which are envisaged to be implemented by the ePortfolio analysis and positioning service are as follows: (1) semantic annotation of the ePortfolio, (2) lexical chains approach to disambiguation of concept annotation (3) discourse annotation of these texts; and (4) sentiment analysis of the discourse segments as well as the mentioned concepts with respect to the levels of learner’s concept competence. They will be combined in a common procedure. The result of the knowledge rich analysis of an ePortfolio will be a concept evidence of the learner’s competence expressed in the ePortfolio. The elements of the concept evidence of the learner’s competence will be a set of concept descriptions extracted from the ePortfolio with links back to the text of the ePortfolio. In this way the concept evidence of the learner’s competence can be automatically compared to a set of concept evidences of the target competence (learning network in the terms of Kalz et al. 2007). Those will be selected that are not covered by the current learner’s competence. For the comparison of the concept evidences we will use the standard vector metrics from Information Retrieval community. The links to the ePortfolio will support the assessors of the student competence to find out the reason for the inclusion of a concept description in the concept evidence of the learner’s competence. The content analysis which is meant to be implemented for this task will allow us to use the methodology for positioning of learners presented in (Kalz et al., 2007). Concept descriptions used for the semantic annotation and for the representation of concept evidences are taken from the domain ontology. Recall that here we consider only an approximation of the learner’s competence based on the concepts from a domain ontology and their usage in the ePortfolio. Much more work will be necessary in order to
support a full representation of the learner’s competence. The same applies to the target competence encoded in the curriculum description.

The semantic annotation and the discourse annotation will be used also in work package 6 (WP6). The difference will be in the domain of application and the specific type of text which will be analysed here, namely the ePortfolio document.

2.5.2.1 Semantic Annotation
In order to use the LT4eL model for the analysis of the ePortfolio we will implement the ontology-to-text relation for the new domain (medicine) with a new vocabulary. We are going to extend the previous implementation with new disambiguation functionality which will be based on lexical chains (Galley and McKeown, 2003), using semantic annotation of general words in the text (in addition to the domain specific terms) and discourse annotation. For the semantic annotation of the general words we will use OntoWordNet (Gangemi et al., 2003) which is already aligned to the same upper ontology which will be used in the construction of the domain ontology.

The output of this new functionality will be a semantically annotated text of the ePortfolio. Each domain term will be annotated with a concept from the domain ontology and each general word will be annotated with concepts from the upper ontology.

2.5.2.2 Discourse Annotation
Similarly to the task within WP6, our main goals in developing an additional layer of discourse segmentation and relations annotation are: (1) to investigate the possibility for refining concept recognition and sense disambiguation for targeted words (lexical terms) via coherence relations (discourse relations, rhetorical relations) markup; (2) the discourse annotation will be used for the sentiment analysis in order to evaluate the learner’s attitude to the concepts mentioned within the ePortfolio. The input for this functionality will be the results from the previous functionality. The discourse annotation process consists of several steps, which may be iterated.

1. Creation of coherence relations taxonomy. We will start with the set of relations and the coding scheme defined by Wolf and Gibson (2006). Their taxonomy is based on the Hobbs list (Hobbs, 1985), and is more coarse-grained than others that include up to 400 types of relations. This fact makes it really applicable for our task. It consists of eleven types of coherence relations:

   **Temporal sequence**: When one discourse segment describes an event that takes place before another event, expressed in another discourse segment.

   **Cause-effect**: When one discourse segment describes the cause, and another – the effect for a given event.

   **Condition**: When one discourse segment describes a possible event that will occur only if another event, described in another discourse segment, also occurs.

   **Elaboration**: When one discourse segment elaborates, i.e. gives more detailed information about another discourse segment.
Example: When a discourse segment provides examples for another discourse segment.

Similarity: When the event, expressed in one discourse segment, is similar to an event, expressed in another discourse segment.

Contrast: When the event, expressed in one discourse segment, contrasts an event, expressed in another discourse segment.

Generalization: When one discourse segment states a generalization for the content of another discourse segment.

Violated expectation: When there is an absence of a causal relation between two discourse segments.

Attribution: When one discourse segment states the source for the content of another discourse segment. It is usually used in constructions, such as: John said that…

Same-segment: Same-segment is a structural type of relation, because it holds between disconnected parts of one discourse segment (subject NP separated from its predicate). Same-segment, similarity and contrast relations are symmetrical while the rest are asymmetrical (directed), that is – one of the segments is more important (the nucleus) than the other (satellite).

According to the coding scheme the three general steps of the annotation process are: (1) the output of the sentence-splitter is segmented further into clauses and then, if needed, annotators insert intrasentential boundaries for smaller discourse segments; (2) the discourse segments are grouped thematically and (3) the coherence relations between the segments are indicated. After the annotation process is finished, the taxonomy may be further adjusted to improve the descriptive adequacy for the texts in the Computer Science domain.


3. The analysis of the obtained discourse structures will provide information that could be used for (1) the development of constraints over the semantic annotation grammar, (2) supporting anaphora resolution, and (3) support of sentiment analysis. In addition, we consider the possibility to create a rule-based grammar for recognizing the coherence relations that are unambiguously linguistically marked.

In order to improve the concept annotation, we will test different knowledge-based techniques that are common for word sense disambiguation. Our main goal is the enrichment of the concept annotation grammar in order to map the relations between text chunks, recognized as carriers of the concepts, with relations present in the domain ontology: is-a, part-of, etc. In the future different algorithms for automatic establishment of lexical chains (with nouns) may be tested (for an overview see Mihalcea, 2007). Lexical chains and rhetorical relations, the two types of discourse information, contributing to the text coherence, will be used for improving the concept annotation. For example, a discourse segment, nucleus in an elaboration relation, will most probably
contain a term, connected via hypernymy relations with lexical units that belong to the
satellite segment.

The discourse annotation will be adapted to the format of the ePortfolio. The idea is that
elements of the ePortfolio will require some specific kind of language. In such cases the
discourse structure might depend on the peculiarities of the corresponding sub-language.

The output of this functionality will be a segmentation of the text of the ePortfolio in
discourse elements and annotation of the relations between them.

2.5.2.3 Sentiment Analysis

The input for this functionality will be the results from the previous above described
functionalities.

In order to construct a concept evidence of the learner’s competence we first need to
extract the concepts which are mentioned within the ePortfolio. Then, on the base of the
ontological reasoning, the implied concepts will be added. For example, if the
ePortfolio’s holder says that he/she is used to giving injections, this automatically means:
on more general level, that he/she can intervene in order to improve the situation, and, on
more specific level, that he/she can put liquid under the skin by using a syringe. We also
need to know in what context each of the concepts in the ePortfolio was mentioned by the
learner. For example, behind the discourse relation, called contrast the learner stated two
opposite fact: it is easy to give an intradermal injection, but it is difficult to give an
intramuscular one. From this short context a conclusion can be drawn that the learner is
not experienced in giving injections as a whole. Thus, comparing conceptual information
and discourse relations, each mentioning of a concept will be evaluated by one of the
values: ‘well known’, ‘known’, and ‘unknown’⁸. We will use the methods developed in
the areas of sentiment and opinion analysis. As it was already mentioned, a pre-defined
requirement list of necessary concepts with definitions will be used in order to estimate
the degree of competence, delivered by the learner in the ePortfolio. There will be three
types of evaluation: coverage, degree of detail and relevance. The coverage will be
estimated over the number of the mentioned relevant concepts that match the pre-defined
list. The degree of detail will be evaluated over the depth of the conceptual space. And
the relevance will be estimated via the ontological relations from a given concept to the
other co-occurring concepts within the discourse segment.

2.5.2.4 Construction of a Concept Evidence of the Learner’s Competence and
Knowledge

As it was mentioned above, a concept evidence of the learner’s competence is a set of
concept descriptions extracted from the ePortfolio. For the moment we divide this set in
the following subsets: (1) known concepts; (2) partially known concepts; (3) unknown
concepts; and (4) concepts with contradictory usages. The first subset will contain all the
concepts which are evaluated as known in the ePortfolio. The second subset will contain
concepts that are mentioned in the ePortfolio, but there is no enough evidence about the
level of knowledge of the learner with respect to them. The third subset will contain

---

⁸ In the process of experiments with the actual data we will refine this scale of values.
concepts that definitely are mentioned as unknown by the learner. In the last subset we will include the concepts for which there are positive and negative evidences about the knowledge of the learner. In addition to the extracted concepts we will extract links to the occurrences of the concepts in the text.

The output of this functionality will be used further to compare the concept evidence of the learner’s competence with the learner network. The comparisons will use a vector representation of concept evidence of the learner’s competence and concept evidence of the target competence. The vector for target competence will be fixed within the learner network. The vector for learner’s competence will be created by the assessor on the basis of the above sets of concepts.

The evaluation of the method will be done on two levels. First, for each of the processing steps we will create manually gold standard corpus on which to test the corresponding technology using the usual precision and recall metrics. Second, we will test the method with respect to the performance of the LSA-based method.

Although we described the method as an alternative to the LSA-based method, we envisage also integration of the two methods. First, in the construction of textmatrix instead of terms from the text the concepts from the conceptual annotation could be used. In this way one can abstract over the textual representation of the concepts. For example, very often in text a super-concept term can be used to denote a sub-concept – “system” instead of “computer system”. Also with the sentiment analysis we could select which concepts to be included in the textmatrix. It is also possible to combine the two methods via combining their results.

2.6 Future Work
In the next cycle of the project, task 4.1 will focus on the design and development of version 1 of the services for ePortfolio analysis and positioning. For supporting the design data sets will be build and validation of (P)LSA and knowledge rich based approaches will be completed. That evaluation includes set up of experiments and the measures of performance (recall, precision, etc) for running a comparative study between both approaches as well as a user oriented study. Based on the exploratory results already obtained, the validation of the showcase will be completed.

Next, the work carried out in the context of WP4 task 2, conceptual development, will be described.

3 Diagnosing Learner’s Conceptual Development
Modern educational approaches stress the importance of collaborative knowledge construction, problem solving, working in groups, personalised and self-directed learning. Despite of their benefits, these activities are highly demanding on time and resources. On the one hand, learners are asked to perform learning activities in groups, using online discussions for develop their knowledge and critical thinking, and providing evidences of their competences; quite often, however, the feedback is given in a social
context rather than on an individual basis. Learners, therefore, find it difficult to realise their individual level of progress.

On the other hand, tutors need to provide appropriate guidance and formative feedback to each learner, but they have to face the issue of a large student population. This might result in limited or ineffective supervision by tutors, which has the potential to result in knowledge gaps remaining undiagnosed as learners fail to engage in tasks too difficult for their level of conceptual understanding or, at the opposite end of the scale, become bored with insufficient challenge and diversity in their learning experience. Tutors, consequently, need to be alerted to the immediate needs of learners that are in need of guidance and should be provided with sufficient information to be able to provide guidance that effectively addresses learner’s knowledge gaps in understanding a topic.

In formal education settings, tutors need to know that learners are developing their conceptual knowledge in line with the expectations laid out in the curriculum, and if this is not the case propose alternative ways to fulfil these expectations. In informal and non-formal settings, in which there is no single learning path to follow, learners need to decide themselves which is the next best step to follow. Receiving formative feedback will help them to make well informed decisions regarding the next best step to take. In both formal and informal contexts, receiving formative feedback will permit learners to be aware of their progress and, consequently, become self-directed learners.

Learners, therefore, must receive formative feedback about how their knowledge of a domain is evolving, in a way that they can determine effectively their advancement and position when they are developing their competences. Tutors require reliable means of analysing the progress of learners, in order to provide appropriate guidance and feedback to each one of them as an individual. In a more general, overview level, other stakeholders, such as curriculum developers, researchers, and so on, might need to have a broad-spectrum view of the conceptual development of learners. All these require determine and diagnose the status of the learner’s conceptual development by eliciting the domain concepts the learner knows, how she relates these concepts to other concepts, and how she uses them, compare them with a reference model, and provide information about this comparison to the learner as an individual, to the tutor regarding groups of learners, and to other stakeholders providing a more general overview as, for instance, regarding groups of tutors, or regarding groups of learners who share special characteristics.

Traditionally, the conceptual development of learners is mostly recorded through the examination and assessment system. In the case of undergraduate medical students these may differ according to individual medical schools but generally the format of the material produced by students is in the form of short responses to questions which address understanding and knowledge or ability to relate this to the clinical context. The former is termed by Schuwirth and van der Vleuten (2004) as Response Formats and the latter as Stimulus Formats. Multiple Choice Questions or Extended Matching Questions are often the format for these assessments and questions may be a combination of Response and Stimulus Formats. Schuwirth and van der Vleuten point out that it is the context of the questions is of major importance. The more complex questions of this type are based on problem solving ability and on the assumption that semantic networks and
illness scripts exist (Schmidt et al., 1996). One of the advantages of the Multiple Choice and Extended Matching formats is that it is highly reliable. One of the disadvantages with this form of assessment, however, is that the use of Multiple Choice or Extended Matching formats limits the opportunity for tutors and clinical teachers to provide feedback and guidance to students on the nature of their conceptual development and the development of learning strategies. To address this, some medical schools have introduced specific formative assessment exercises (Krasne et al., 2006). Although these are viable in smaller medical schools, they require extensive tutor time if they are to be applied at different times in the academic year, to chart students’ development. The ability to self assess is also recognised as important, especially as medical schools are required to prepare students as independent lifelong learners. Although students can assess their development on the large scale and can judge their performance after assessment, their expectations of their future assessments were far less accurate and realistic (Eva et al., 2004). A means of providing students and tutors with clear understanding of individual students’ conceptual development, which is economical with tutor’s time, is therefore required.

In the Language Technologies for Lifelong Learning project, we hold that the kind of feedback described before is part of a new generation of support and advice services needed to enhance individual and collaborative building of competences and knowledge creation. Regarding learner’s study progress, the project will offer continuous modelling and measurement of conceptual development. The final outcomes of task 4.2 are services and tools that determine and diagnose the status of the learner’s conceptual development in order to identify knowledge gaps and recommend remedial actions. In consequence, this first part of the project focuses on exploring until what extent Language Technologies can be used to monitoring the conceptual development of the learners.

Research on the application of Language Technologies (LTs) indicates that it is possible to measure conceptual development both with regard to reliability and validity (Clariana and Wallace, 2007; Koul et al., 2005). Our focus is, however, to demonstrate if, in certain learning scenarios, LTs perform equally well or sufficiently better on diagnosing and monitoring learner’s conceptual development than other methods and tools such as, for instance, the concept map method and existing tools to create concept maps tools – as KNOT (Clariana et al., 2006) or those described in (Shute et al., submitted) –, or semantic networks to represent knowledge and distinguish between levels of expertise – as those described in (Budé, 2007; Schmidt and Boshuizen, 1993; van de Wiel et al., 2000); or if the result of this comparison is negative, whether the derived LTs solution might be in combination with existing methods to improve the solution.

The research question that could be derived from this is:

*To what extent can Language Technologies (e.g., Latent Semantic Analysis, Natural Language Processing, etc.) contribute to measuring learners’ conceptual development when compared to or combined with other methods and tools that measure conceptual development?*

By directly using language technologies to analyse the learner’s conceptual development, our purpose is to either reduce the constraints imposed by existing methods and tools that aim at conceptual development (e.g., the expertise and time required to use them, pre and
post-tasks required, etc.), and therewith extend the number of situations in which learners can obtain feedback on their conceptual development, or to enrich the feedback possible by combining existing methods with LTs.

The outcomes of this exploration will give requirements to develop services that diagnose learner’s conceptual development, inform them of possible knowledge gaps and recommends remedial actions. Moreover, the services will provide information to tutors and other stakeholders so they can monitor learner’s conceptual development.

Although the services to be developed are intended to be used in all knowledge domains, the starting point is the medical education domain. The rest of this section draws up first on the theoretical foundations about conceptual development, which focus mainly on medical education, and explains the process that needs to be followed for measuring conceptual development. Next, the approaches that have been defined to check the sufficiency of LTs against existing tools and methods are presented, along with the initial explorations already done.

3.1 Theoretical Background

In diagnosing conceptual development one has to cope with interrelated qualitative changes that occur and turn learning into a stage-like process, in which each stage is characterized by different learning processes and different effects on knowledge structure. These differences are due to structural changes in the knowledge base development from novice to expert (Boshuizen and Schmidt, 1992). Therefore, diagnosing learner’s conceptual development could be performed by comparing the knowledge of a learner with the knowledge an expert-learner would have, or it is required to have, in a particular context. Following this idea, the rest of this section elaborates on the distinction between novice, intermediate and experts, and the implications of diagnosing medical students’ conceptual development. Then, the section presents the components and characteristics of the “structural approach” (Goldsmith et al., 1991), a process to measure conceptual development.

3.1.1 Learners as Experts, Intermediate or Novices (in Medicine)

Aristotle said the expert “straightway” does ‘the appropriate thing, at the appropriate time, in the appropriate way’ (Dreyfus and Dreyfus, 2005). The definition of the term expert, nevertheless, is not straightforward.

Davenport and Prusak (2000) mention that "experience" and "expert" are related words, both derived from a Latin verb meaning “to put to the test.” Experts -people with deep knowledge of a subject- have been tested and trained by experience. An expert can be also defined as a top performer who excels in a particular field such as arts or athletes. In research on expertise, and particularly in the context of conceptual development, an expert is defined as a professional who achieve certain level of success in her occupation (Boshuizen et al., 2004). For these experts knowledge includes not only declarative and procedural knowledge, but also attitudes, and the so-called enculturation, in which professionals acquire skills, attitudes and habits of a certain profession and become accepted and legitimized in certain context.
According to Arts et al. (2006), research on expertise, in areas such as Management and Medicine, has shown that experts and novice differ in their problem-solving skills, knowledge use, information processing, time required for diagnosing, and on the organization of their knowledge structures. Experts make more appropriate diagnoses than novices, provide more accurate problem solutions and make less use of theoretical knowledge during problem solving than novices. Furthermore, experts use less time to provide diagnosis and distinguish better between relevant and non-relevant information than students that tend to reason on both relevant and irrelevant information (Boshuizen and Schmidt, 2004).

In addition, experts have elaborated, well structured and organized mental frameworks that they activate to interpret information and problems and create a suitable solution (Boshuizen and Schmidt, 1992; van de Wiel, 1996). In contrast, novices do not easily activate their mental frameworks which are, furthermore, less accurate, completed, organized and structured (Nievelstein et al., in press). In fact, Nievelstein et al., (Nievelstein et al., in press) have found that, in the private law domain, knowledge becomes more hierarchically structured as expertise grows. Novices’ knowledge appears to be highly fragmented and concepts loosely connected. These findings correspond to those of expertise research in other domains as, for instance, Physics (Dufresne et al., 1992), Management (Arts et al., 2006), and Medicine (van de Wiel, 1996; van de Wiel et al., 2000).

It is important to notice that the differences between experts and novices in professions in which diagnosis (e.g., medicine, chess, engineering, statistics) is the central task, cannot be explained by their reasoning skills, since both experts and novices reason by making a diagnosis based on a hypothesis. The difference is the accuracy of the diagnosis. Studies on the development of medical expertise (Boshuizen, 2003) reveal that to formulate a diagnosis medical students, who have not yet gained practical experience, focus much more in biomedical knowledge (i.e., knowledge about functioning and malfunctioning of the human body) than in clinical knowledge (i.e., knowledge about the classification and treatment of diseases). Once medical students have gained practical experience their biomedical knowledge usage decreases. Nevertheless, physicians use as much biomedical as clinical knowledge to formulate a diagnosis (van de Wiel, 1996).

The professional learning process in medicine has been split by Boshuizen and colleagues (Boshuizen et al., 2004; Boshuizen and Schmidt, 1992) into three levels: knowledge accreditation, knowledge encapsulation and illness script formation. Each one of them corresponds to an expertise level: novice, intermediate and expert/experienced expert. Table 3 shows how (Boshuizen et al., 2004; Nievelstein, 2004) summarize the knowledge structure, learning, problem solving, reasoning process and demand on cognitive capacity of these levels.

Summarizing the table (for a detail description see (Boshuizen et al., 2004; Nievelstein, 2004), novices structure knowledge in networks, which represent small steps of reasoning with self-explanations. They use every day knowledge terms and have deeply linked integrated knowledge; when they try to make longer chains, they are less self assured. Novices rely only on knowledge networks that are less rich and less easily activated than illness scripts, which represent the patterns experts use to describe the process of
contracting a disease. To make a diagnosis or solve a problem, novices require more information and, as a result, semantic networks must be reasoned step-by-step. Their reasoning is less ordered, less goal-oriented, more time consuming and it is based on less plausible hypotheses resulting in less accurate diagnoses than those from experts.

Intermediate students increase the step size by gather together multitude of detailed concepts ‘encapsulated’ under one higher order concept. Boshuizen and Schmidt (1992) report a model of expertise development in medicine towards knowledge encapsulation, which is defined as the acquisition of biomedical knowledge, practical experience, and the integration of theoretical experimental knowledge.

Finally, experts use illness scripts to structure knowledge. When they deal with a case they activate ready-made illness scripts as a whole, which means no small steps between them are taken. This activation depends on information about the conditions (i.e., factors that affect health in positive or negative way), fault (i.e., path-physiological process taking place in disease), and consequences (i.e., signs and symptoms of a specific disease) (Boshuizen and Schmidt, 2004).

3.1.2 Diagnosing Conceptual Development: The Case of Medical Schools

Conceptual development in medical schools is highly influenced, widely throughout Europe, by the problem based learning (PBL) approach. This constructivist approach emphasises the experiential nature of learning and challenges the division between “theoretical” learning (which used to be confined to what was termed “preclinical years”) and its application to the practical work place context of the “clinical” years. This is reflected by the introduction of Early Experience, whereby students have contact with clinical practice from the outset of the curriculum, and of problem based learning.

Although these developments were introduced in response to examples of poor medical practice, which raised issues concerning the education of clinicians, they are grounded in modern educational theory. Most important is Situated Learning Theory (Lave and Wenger, 1991) and in Wenger’s approach of a community of learners, which acknowledges and develops further the “social” component, intrinsic to cognitive development (Wenger, 1998). In the modern medical school, students experience several types of Communities of Practice. Apart from the “enculturation” mentioned earlier, students experience learning in a group setting through their problem based learning. Here knowledge is shared and debated and the understanding obtained by individual group members is tested against that which has been obtained by others in the group. The aim is that the understanding obtained by the group as a whole (i.e. the communities of learners) will exceed that of the individual members.

The importance of problem based learning, however, is not confined to the interactions within the group, as developing communities of practice and learning, but is realised through the method of analysing problems, which leads to the student reaching his/her intended learning objectives. The approach, described by O’Neill et al. (2002) requires students to select “cues” or triggers from a case, link them together into a hypothesis, from which learning objectives are identified and to test this first through individual study and then through group discussion. This broadly corresponds to “clinical
reasoning”, mentioned earlier and is the main driver for learning throughout the curriculum. The consequences of this are that the students are more familiar with an integration of systems from the basic sub cell, on the micro level to the psychological and social implications of illness and disease on the macro level.

<table>
<thead>
<tr>
<th>Expertise Level</th>
<th>Knowledge Structure</th>
<th>Learning</th>
<th>Problem solving</th>
<th>Reasoning process</th>
<th>Demand on cognitive capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Novice</strong></td>
<td>Networks (incomplete and loosely linked)</td>
<td>Knowledge accretion, integration and validation</td>
<td>Long chains of detailed reasoning steps through networks</td>
<td>Step by step process</td>
<td>High</td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td>Networks (tightly linked and integrated)</td>
<td>Encapsulation</td>
<td>Reasoning through encapsulated network; abbreviated</td>
<td>Big steps (but still one at the time)</td>
<td>Medium</td>
</tr>
<tr>
<td><strong>Expert</strong></td>
<td>Illness scripts</td>
<td>Illness script for formation</td>
<td>Illness script activation and instantiation</td>
<td>Groups of steps activated as a whole</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Experienced expert</strong></td>
<td>Memory traces of previous cases</td>
<td>Instantiated scripts</td>
<td>Automatic reminding</td>
<td></td>
<td>Low</td>
</tr>
</tbody>
</table>

Table 3. Expertise level, knowledge structure and learning

It is important to notice, therefore, that an approach as the problem based learning, which requires students to learn in communities of practice, and to employ clinical reasoning as a foundation for their learning and assessment, the demarcation of the stages of conceptual development are likely to be less clear cut than in a curriculum which is based on acquiring knowledge first and then applying it to a work related situation. This implies that a rigid distinction between novices and experts may not be applicable in all circumstances. Such a distinction, however, should be considered as a useful to provide information to each student about her current level of expertise compared with the level of expertise required in a particular level or course, and to provide feedback based on this comparison. **The “expert”, therefore, is considered as someone that has the required level of conceptual development for a particular context.**

From our point of view, problem based learning does have some drawbacks, which are mainly related on the practical implementation of this approach, rather than on its principles. First, as students are expected to develop their knowledge in communities of practice and in groups, often it is difficult for them to realize their individual level of conceptual development and their progress during the course. Second, providing feedback is a time consuming tasks for tutors and, frequently, it is only provided at a group level, as a consequence students might do not receive appropriate, on demand, individual feedback.

Regardless the educational approach, the question is how medical students can be supported so they get formative feedback which might include their current level of
experts in a particular context, the difference between their level and the level required in such a context (the “expert level”), an identification of their knowledge gaps, and recommendations of remedial actions. To this end, an important contribution can be given by measuring the learner’s conceptual development.

### 3.1.3 The Process of Diagnosing Conceptual Development

An well known example on how the conceptual development can be measured is the structural approach proposed by (Goldsmith et al., 1991), which assess the person’s knowledge of a domain by looking on how she organises the concepts of such domain. This approach involves three steps: knowledge elicitation, knowledge representation, and evaluation of an individual’s knowledge representation. In the rest of this section we describe these three steps.

#### 3.1.3.1 Knowledge Elicitation

Knowledge elicitation is defined as the process of describing domain specific knowledge underlying human performance (Cooke, 1999). For a review of knowledge elicitation techniques see (Cooke, 1994). In short, knowledge elicitation techniques measure the person’s understanding of the relationships among a set of concepts (Jonassen et al., 1993). Methods that support this activity include categorization, graphical representation and verbal reporting methods. Table 4 shows some examples of these methods.

<table>
<thead>
<tr>
<th>Categorisation methods</th>
<th>Graphical representation methods</th>
<th>Verbal reporting methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card sorting</td>
<td>Concept maps</td>
<td>Thinking-aloud</td>
</tr>
<tr>
<td>Word association</td>
<td>Semantic networks</td>
<td>Essay questions</td>
</tr>
</tbody>
</table>

Table 4. Examples of knowledge elicitation methods

Cooke (1994) claims that the suitability of each method depends on the type of knowledge to be elicited. For instance, card sorting is more appropriate for declarative knowledge, whereas thinking aloud protocols are more appropriate for procedural knowledge. Cooke also stresses that, in general, no technique can ensure the complete and accurate representation of the knowledge. She also points out that the main back of knowledge elicitation methods is that they can be costly: get rich data implies long data collection sessions, cumbersome data analysis, and interpretation difficulties.

**a) Categorisation methods**

**Card sorting.** The aim of this a technique is to understand how a person classifies and relates items. Items are printed in cards and participants are asked to group items into piles, which represent categories. There are different card sorting techniques. The simpler one is to ask participants to sort cards into piles in any way that makes sense to them, but without either placing the same card in more than one pile or placing all the cards in only one single pile (Trochim, 1989). Other techniques include group separation tasks, in which participants are asked to sort cards in two groups that the participant should name,
and group creation tasks, in which participants are asked to find a pair or cards that are more similar than any other possible pair (Wright and Ayton, 1987).

Card sorting methods provide information about differences on the organisation of conceptual knowledge amongst individuals with different levels of expertise (Nievelstein et al., in press); they do not, however, provide information regarding the kind of relationship between the concepts. It is useful, therefore, to perform a follow-up activity in which participants explain why they grouped concepts in certain categories and how each item differs from the others in a category (Welbank, 1990).

**Word association.** The aim of this technique is to elicit through recall information about the structural interrelatedness of ideas. For each concept in a domain, the participant is asked to generate a list of associated words that immediately come to her mind and free associate each one of the concepts. An alternative form is using a controlled word association procedure in which the participant ranks the words as she associates them (Jonassen et al., 1993).

### b) Graphical representation methods

**Concept maps.** Originally developed by Novak (1977), concept maps are a mean to describe and visualize graphically how a person thinks concepts of a particular domain are related. The technique is meant to externalise learners’ mental models in an intuitive and flexible way by applying a simple graphical format: nodes represent concepts and labelled links their relationships. Concepts are connected by arcs which contain propositions about the type of relationship (“is part of”, “belongs to”). In this way, concept mapping is a concise and parsimonious technique, which is at the same time rich in information, because of the integration of verbal and visual coding. As a knowledge representation tool, concept maps have been used as learning or metacognitive tools or as an evaluation tools (Novak, 1998).

According to Cañas et al. (1997) concept maps that use verbs are more likely to be richer in terms of explanations and represent deeper thoughts. In contrast, maps made mainly with objects are more descriptive. Similarly, Jonassen et al. (1993) argue experts present problems in terms of abstract principles, whereas novices present problems in terms of literal characteristics.

**Semantic networks.** Represent semantic relations between the concepts. They are directed or undirected graphs consisting of vertices, which represent concepts, and edges, which represent the relations between the concepts (Sowa, 1987). The main difference between semantic networks and concept maps is the arrangement of the central concept and the related concepts. Concept maps represent concept in a hierarchical structure, which places the more general concepts on the top and more specific ones at lower level, whereas semantic networks place the concept is at the centre and the related concepts around in a radial fashion way.
c) Verbal reporting methods

**Think aloud.** The main purpose of this technique is to obtain information that normally is not available through other methods (Ericsson, 2006). In thinking aloud sessions participants are asked to verbalise their thoughts during the performance of a task or problem-solving. If necessary, this request is repeated to encourage the person to tell what she is thinking.

This method is used to investigate differences in problem solving, differences in difficulty between tasks, effects of instruction or other factors that might have an effect on problem solving. Compared with other elicitation methods, the think aloud method is considered easier, as it encourages people to use their own language, the structure of the knowledge is done by the person who analyses the protocols (van Someren *et al.*, 1994).

**Essay questions.** One of the most used methods to assess learner’s knowledge, essay questions aim is to elicit learner’s knowledge by asking them to compose a text that answers a particular question.

3.1.3.2 Knowledge Representation

The second step of the process is to define some representations of the elicited knowledge that reflect the underlying organization of the data (Goldsmith *et al.*, 1991). To this end, advanced statistical methods (e.g., cluster analysis, tree constructions, dimensional representations, pathfinder nets) are used to discover the structural framework underlying the set of concepts (Jonassen *et al.*, 1993).

In some cases, eliciting and representing knowledge are performed together. While in others, knowledge representation is doing by pre-processing the elicitation knowledge, structuring and adding useful information, in such a way that it can be evaluated in the next step of the process.

3.1.3.3 Evaluation of the Representation

The third step is to evaluate the individual’s knowledge representation relative to some standard (e.g., expert’s organisation of the concepts in the domain, reference model, etc.). Normally, researchers follow one of these three approaches (Goldsmith *et al.*, 1991):

- Qualitative assessment of derived representations
- Quantifying the similarities between a student representation and a derived structure of the content of the domain
- Compare the cognitive structures of experts and novices

Interestingly, semantic networks have been used to represent knowledge and compare cognitive structures of experts and novices (Budé, 2007; Schmidt and Boshuizen, 1993; van de Wiel *et al.*, 2000). For instance, van de Wiel *et al.* (2000) conducted an

---

*Note, however, that a huge range of different computations for representing knowledge from concept/network maps that can be used. It is still an open question if new evaluation protocols should be developed or an existing one could be used.*
experiment to investigate the qualitative changes that occur in the structure of knowledge in acquiring medical expertise. They asked medical students and physicians to provide diagnoses of some clinical cases, providing the pathophysiology underlying the signs and symptoms. To assess the accuracy of the diagnoses the explanations provided by students and physicians were manually rewritten as semantic networks of meaningful concepts. Using a classification developed by (Schmidt and Boshuizen, 1993), the links were qualified as: causal, conditional, temporal, attributional, locational, specification, negation, identity relation, a class relation. To identify the differences between knowledge encapsulation in novices and experts, several parameters were defined as, for example:

- Total number of different concepts applied (correct concepts and few concepts = expert)
- Quality of the explanation = % of concepts/total concepts
- Number of detail concepts (lower the number of detail concepts = higher encapsulation)
- Total links between concepts (higher the links between the concepts = better explanation)
- Shortcuts (i.e., skipping intermediate nodes), (more shortcuts = expert)
- Number of alternative links (low level of alternative links = expert)

Likewise, semantic networks were evaluated by Budé (2007) in order to put into practice a procedure to measure learner’s understanding of a science text. The evaluation took into account:

- The total number of mentioned concepts
- The total number of mentioned links
- The number of concepts matching the model concepts
- The number of links matching the model links
- The number of extra correct concepts
- The number of extra correct links
- The number of shortcuts due to concise but correct answers
- The number of all possible model concepts
- The number of all possible model links

The accuracy of the answers was determined by two ratios:

\[ R1 = \frac{(C + E)}{A} \]
\[ R2 = \frac{(D + F + G)}{B} \]
The completeness of the answers was determined by two other ratios.

\[
R3 = \frac{C + E}{H}
\]

\[
R4 = \frac{D + F + G}{I}
\]

The four ratios were considered complementary in establishing the quality of the participants’ answers. They expressed how much of the answers were correct and how complete the answers were.

### 3.2 Description of the Work to Be Done

In task 4.2 we will use language technology software applications to identify/approximate the conceptual development of learners. For this purpose we will use and evaluate several existing software products and evaluate how well they are suited for monitoring conceptual development of learners. One approach will focus concept maps while the other approach will be solely focused on Latent Semantic Analysis. Both approaches will follow the three steps described earlier: knowledge elicitation, knowledge representation and evaluation of the individual’s knowledge.

For the knowledge elicitation we will use verbal reporting methods. At the first stage a combination of a structured questionnaire with a think aloud session will be used. The topic of the elicitation will be a specific PBL-case (c.f. Appendix 8). For the knowledge representation and analysis a combination of LSA and different analysis techniques will be investigated. In this section we explain our approach in more detail. In the following section we will explain our two main strands of research more into detail. Appendix 1 (Task 4.2) describes a showcase that illustrates how these two approaches will be implemented in a first pre-pilot scenario.

#### 3.2.1 Existing Methods and Tools Approach

It has been decided to start the exploration of the cognitive map method, which is one of the most common methods for representing cognitive structures, as a mean to elicit and represent learner’s knowledge. The decision was taken on the basis of the appropriateness of concept maps for representing learners’ representations of subject matter structure and on research evidence that points out concept map method as well suited for eliciting knowledge (Nesbit and Adesope, 2006), and as better method for evaluating meaningful learning of learners of different ages than classical assessment methods such as tests and essays (Jonassen et al., 1997; Novak, 1998). It is important to point out, however, that the creation of concept maps is a complex and time consuming task that requires training and practice to understand how the relevant concepts should be identified and how to make the relation between them.

There are already a number of tools for automatic construction and support of the construction of concept maps: Knowledge Network and Orientation (KNOT, PFNET) (Clariana et al., 2006); Surface, Matching and Deep Structure (SMD) (Ifenhaler and Seel, 2005); Model Inspection Trace of Concepts and Relations (MITOCAR) (Pirnay-Dummer, 2006); Dynamic Evaluation of Enhanced Problem Solving (DEEP) (Spector
and Koszalka, 2004); jMap (Jeong, 2008), and ProDaX (Oberholzer et al., 2008) Table 5 depicts these tools in terms of the data collection they use, the analysis they perform, the data conversion they use and the comparison they perform.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Data Collection</th>
<th>Analysis</th>
<th>Data Conversion</th>
<th>Comparison(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNOT</td>
<td>Concept pairs/Propositions</td>
<td>Quantitative Analysis</td>
<td>Pathfinder Networks</td>
<td>Direct comparison of networks with some statistical results.</td>
</tr>
<tr>
<td>SMD</td>
<td>Concept map or natural language</td>
<td>Quantitative—analysis is calculated using tools.</td>
<td>Structural decomposition into 3 categories (manual and semi-automatic)</td>
<td>Unlimited comparison</td>
</tr>
<tr>
<td>MITOCAR</td>
<td>Natural language</td>
<td>Quantitative—analysis included multiple calculations using tools</td>
<td>Structural composition into 1 category (automatic)</td>
<td>Paired comparisons for semantic and structural model distance measure</td>
</tr>
<tr>
<td>DEEP</td>
<td>Annotated causal maps</td>
<td>Quantitative/qualitative analysis — analysis is done mostly by hand</td>
<td>Structure decomposition into 3 categories (automatic)</td>
<td>Unlimited comparisons, showing details relative to concepts</td>
</tr>
<tr>
<td>jMap</td>
<td>Concept maps, causal maps, or belief networks</td>
<td>Quantitative analysis – analysis is calculated using tools</td>
<td>Structural decomposition into link strengths between causal factors and evidentiary strength</td>
<td>Superimposes maps of individual (n=1) and group of learners (n = 2+) over a specified target map</td>
</tr>
<tr>
<td>ProDaX</td>
<td>Association Data, Cross-Tables, Two-Way Two-Mode Data, Coordinates, Scales</td>
<td>Non-Metric Multidimensional Scaling/Cluster-Analysis</td>
<td>Concept Maps</td>
<td>Comparison of maps based on Procrustean Transformation/Loss-oriented Meta Map (LOMM)</td>
</tr>
</tbody>
</table>

Table 5. Overview of concept mapping tools (adapted from Shute et al., submitted)

These tools have some common characteristics: (a) they are concerned with conceptual development of learners; (b) they can (semi-)automatically construct concept maps from a text; (c) they use a sort of distance matrices; (e) they propose a quantitative analysis of the maps; and (d) most of them pretend to support high levels of learning, namely critical thinking and problem solving.
Amongst their differences we have found that, even though, they all use some sort of Language Technology analysis, not all of them refer to it explicitly. The SMD and jMap can use as an input not only text but also concept maps. These tools also differ on the scoring schemas they use to perform the quantitative analysis: DEEP uses a number of nodes and links; SMD uses propositions or a number of the links of the shortest path between the most distant nodes.

Most of the referred concept mapping tools provide opportunities to identify the conceptual gap between a learner’s concept map and a criterion map (in fact, an expert map), or to compare a learner’s concept maps in different periods of time. However, only SMD, jMap and, in some extent DEEP, provide purposely a visualisation of this progression towards the criterion. Most of these mapping approaches construct and analyse individual maps. jMap visualises and assesses changes observed in either individual or collective maps. Nevertheless, jMap is the only tool restricted to produce a particular type of maps, causal maps.

KNOT, SMD and MITOCAR do report on reliability and validity criteria correlating, as a typical case, the automatic scores generated by these concept mapping approaches and human concept mapping scores and human essay scores. Finally, it is worth to mention that SMD and MITOCAR report experimental data on the effectiveness of a particular technique as an increase in similarity between a learner’s map and an expert’s map.

### 3.2.1.1 Concept Map Approach

A first exploration of existing tools that create concept maps from an input test was performed. The aim was to investigate in which and to which extent existing tools support the process of knowledge elicitation to evaluation. More in particular to gain insight in how flexible and easy to use the tools are (other aspects such as reliability and validity have been derived from literature).

The following tools have been explored:

- CMAP (Institute for Human and Machine Cognition, 2004)
- KNOT and Ala-reader (Clariana et al., 2006)
- INFOMAP (Peters, 2005)

KNOT is a software tool that generates text propositions files that can be imported from CMAP to generate automatically concept maps. Appendix 4 shows the results of this initial exploration. The conclusion of this exploration was that the process could be used to analyze conceptual development but there are restrictions on the data that can be used like, for instance, the general limit of concept pairs that can be used.

Next, an initial exploration of INFOMAP to generate an associative semantic network, based on learner’s texts, was performed. INFOMAP employs a similar approach as Latent Semantic Analysis with a focus on word-to-word relations and a context limitation around the words used for indexing.
To this end, a data set from a Psychology course of the OUNL was used. This data offers course content, which was considered as the expert level of argumentation\(^\text{10}\). Also documents written by the learners were used. For every part of the course content some related keywords were generated, then the inter-correlation of the keywords have been calculated for one exemplary chapter (for details see Appendix 3). After that by using a clustering method (nearest neighbour approach), a distance matrix and clusters in the keywords were generated. Figure 3 shows an exemplary cluster overview of a chapter. These keyword clusters can be used to identify topical foci of the documents. An alternative approach to use the keywords and associated other concepts in documents is Multidimensional Scaling (MDS). With this approach distances between concepts can be visualized (see Figure 4).

\(^{10}\) Notice that in this first exploration “expert knowledge” has been defined as the course content. The “expert knowledge” can be also seen as the knowledge a learner, who is considered by the tutor as an expert learner, has in a particular context.
As a next step highly related associated concepts to these keywords were added to this representation. Again, INFOMAP was used to calculate these associated concepts. Giving the fact that the learner’s text have been provided from several contexts, rather than from only one study context, and since similar words for the concepts were not taken into account, the dataset used was not ideal for this kind of analysis. While the original keywords were not found often in the learner’s document related concepts could be found in these documents and it needs to be evaluated if this approach can in the student documents as a proxy to these important concepts. As another alternative method we will apply Non-Metric Multidimensional Scaling (NMDS), which is integrated in the ProDaX software environment (Oberholzer et al., 2008). Based on generated concept maps from experts and learners ProDax allows a comparison through the procrustean transformation method.

As an alternative approach the use of natural language in combination with Latent Semantic Analysis is a promising step towards the monitoring of conceptual development of learners. This approach will be described next.

**3.2.2 Latent Semantic Approach**

The second approach is focusing on the investigation of unstructured produced text using LSA based methods. For measuring conceptual development in a first experiment...
students’ writings where compared against existing model solutions. Conceptual development of learners is reflected in the texts they are writing. Through structure and word-choice, most notably the application of professional language, their arrangement and meaning give cues about the level of competence development.

When trying to evaluate those writings using LSA in order to monitor the conceptual development, several issues have to be addressed:

- First, the analysis of performance from behaviour can always only provide limited evidence into the competence of the person under investigation. Competence is a potential for action (cf. Erpenbeck, 2003; Erpenbeck and Rosenstiel, 2003; Fiedler and Kieslinger, 2006; McClelland, 1973)– not the action. Any writing therefore can only provide a cue about how competent somebody is. When there are more cues provided, i.e., when there is more evidence, predictions will get better. They, however, will always remain predictions.

- Second, not necessarily all concepts in a professional language are directly realized in language. Research on cognition gives reasonable evidence that there are at least two separate systems for sensory and for semantic (i.e. conceptual) encoding (cf. Engelkamp, 1998; Levie, 1987; Nelson, 1979; Paivio, 1978, 1986; Wild, 2004). The binding of concepts to language happens in the corresponding sensory subsystems, relating words with their meanings in the higher-order, conceptual parts. There are, however, also concepts that are reflected visually, think of, for example, a geometric proof in mathematics or a RGB colour space in design theory. Some concepts even do not have a direct sensory correlate (like emotions).

- Third, the identification of representative textual material to evaluate writings against is cumbersome, sometimes even impossible. Often there is no widely agreed curriculum available, especially not on this conceptual level of detail. Moreover, the identified reference material might differ in wording and writing style compared to the records of action brought forward as evidence: a text book uses different vocabulary and different rhetorical structure than a work report. When striving to find textual model solutions, these may be difficult to get.

- Fourth and related to this, scalability is problematic. The preparation of a reference corpus and the training of the algorithm may be to work intensive to pay back in the little time the conceptual knowledge covered by it is relevant. Training and validation efforts have to be way below the turnover time of the knowledge under investigation. Last but not least, the unit of analysis is not easy to find. Concepts in professional language exist on extremely varying scales, which makes it difficult to define the granularity of the evaluation.

- Fifth, the unit of analysis is not easy to find. Concepts in professional language exist on extremely varying scales, which makes it difficult to define the granularity of the evaluation affecting both the size of the evidence and reference material to be brought forward.

- Sixth, LSA needs threshold values for text categorization purposes. Traditionally the threshold is obtained by calculating the average similarity between texts that correspond to the same category. This procedure can be inaccurate if a representative set of documents for each category is not available. Furthermore, similarity values
tend to decrease with increasing corpora and vocabulary sizes. Building data sets can be costly and there is no warranty that outliers are excluded.

One method to evaluate texts regarding their conceptual coverage is by comparing them against a structured collection of ‘concept-sized’ model solutions. These expert generated model solutions serve as a kind of ‘gold standard’ against which learners’ contributions are evaluated. The model solutions are best-practice essays, typically written by a lecturer or by former students, each covering a particular topic a knowledgeable expert in this subject, context, and point in development has to be acquainted with. Here we will explore the approximation using a final gold standard text and the use of several gold standard texts, which represent conceptual development stages.

Another possible approach to evaluate conceptual development from unstructured texts is to compare the conceptual coverage of a learner’s writings in two points in time. In between those two points, learning has taken place, i.e. a change in the conceptual representations internalized by the learner has taken place. This change may take different forms: it can extend the concept space (or remove misconceptions) or it can change their organisation by rearranging and relating. In its simplest form, the extension of the concept space should be reflected by the writings in so far as new concepts and more concise concepts should be covered compared to the point in time before the learning activities took place. Again a gold standard expert text is used here.

By comparing the similarity of the learner’s writing with the model solution, those concepts can be identified as being covered which match to the input text. By comparing a collection of one learner’s writings and the structured collection of model solutions (structured along the conceptual space to be covered by a knowledgeable expert), insight in the current standing of the learner can be gained. And by comparing the writings around two points in time against the same collection of model solutions, conceptual development can be measured.

For this approach it is assumed that the fewer the divergence between the student’s writing and the model solution is, the higher the correlation should be, thus reflecting that the student covers a specific topic in an adequate sense. The other way round it is expected that a low correlation should mirror a lack of knowledge in this case.

In a first test experiment, students’ writings were evaluated against nine model solutions. To ensure that the model solutions were reflected correctly in the higher-order latent-semantic space, the space was calculated not only from generic background text material but additionally from a considerate amount of representative domain-specific text material that cover the desired concept space of the target domain.

In order to investigate both the concept space covered by the collection of model solutions and compare it with the concept space covered by a student’s contributions, the recombined, factor-truncated text matrix $M_k$ is generated by an LSA process (see equation 1). The learner contributions are added ex post to keep them from influencing the space. Similarly, they generate a resulting text matrix $E_k$ (reusing the left-hand side $T_k$ and the diagonal matrix $S_k$ of the original decomposition).

\[(1) \quad M_k = T_k S_k D_k^T\]
When evaluating one contribution against one model solution, cosine is used to measure the angle between the two representing document vectors (see equation 2)

\[
(2) \quad \text{similarity} = \cos(E_k^n, M_i^j)
\]

The validity of this experiment is measured by comparing a greater number of student essays against each model solution and by comparing the machine assignments with human expert ratings. As a test, spearman’s rank correlation test was applied.

<table>
<thead>
<tr>
<th>essay</th>
<th>Rho</th>
<th>p-value</th>
<th>result #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.42</td>
<td>0.000</td>
<td>5234</td>
</tr>
<tr>
<td>2</td>
<td>0.60</td>
<td>0.000</td>
<td>6759</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0.000</td>
<td>9386</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.000</td>
<td>3543</td>
</tr>
<tr>
<td>6</td>
<td>0.61</td>
<td>0.002</td>
<td>7347</td>
</tr>
<tr>
<td>7</td>
<td>0.58</td>
<td>0.000</td>
<td>3560</td>
</tr>
<tr>
<td>8</td>
<td>0.64</td>
<td>0.000</td>
<td>8167</td>
</tr>
<tr>
<td>9</td>
<td>0.56</td>
<td>0.000</td>
<td>8235</td>
</tr>
<tr>
<td>10</td>
<td>0.71</td>
<td>0.000</td>
<td>10597</td>
</tr>
</tbody>
</table>

Table 6. Optimised evaluation of LSA-based scores against human scores

As can be seen from Table 6, most of the collections could be evaluated with a high rank correlation coefficient while at the same time being highly significant on a level below 0.001 or 0.005. The evaluation is selected from a random sample of 100,000 test-runs (from over 1.3 million) of changing parameter settings for the LSA configuration.

With this very early experiment it is wanted to be proven that using LSA methods on free-texts (student essays) can automatically judge the quality of the writings in-line with human assessments. This is a requirement for computationally measure conceptual development and positioning of a learner.

### 3.3 Future Work

In the next cycle of the project, task 4.2 will concentrate on the implementation of the pre-pilot described in this document (see Appendix 1). That includes data collection as well as the development and configuration for the unstructured NLP approach and existing methods/tools approach. After this, an evaluation will be conducted. It will include an evaluation of the data collection in terms of sufficiency and feasibility, and a comparison between the results of the two approaches. This will result in a set of requirements for developing the first version of the services for diagnosing and monitoring conceptual development and also will point out new needs to explore further different technical and pedagogical aspects.

Based on this, the next step will be to develop the version 1 of the services. In parallel the preparation for pilot 1 will be conducted. This preparation includes describing
educational scenarios in terms of the solution, problem information and interaction scenario. Particular emphasis will be given what and in which way the feedback resulting from the diagnosis will be presented to the users. Considering these, then, the description of the knowledge elicitation procedure (data collection, protocols, etc.), knowledge representation and evaluation of the representation for the pilot will be done, as well as the description of the validation design. Afterwards, the pilot will be run, and its evaluation and validation will be conducted.

4 Conclusions

This report presented the work that has been carried out, from the theoretical, conceptual and exploratory point of view, to decide on how learners’ prior knowledge and conceptual development should/can be identified. The positioning of an individual learner results in a collection of suitable learning materials that will facilitate the achievement of the learner’s learning goals, whereas the diagnosis of learner’s conceptual development will provide learners with formative feedback and recommendations of remedial actions to reduce their knowledge gaps.

The report also described of a service for ePortfolio analysis and positioning, and a plan to extend available infrastructure for LSA and knowledge rich based analysis. Regarding the diagnosis of conceptual development, this document depicted the two approaches that will be conducted to evaluate the differences between using existing methods and tools or “pure” LSA to diagnosis conceptual development.

Furthermore, this report showed some preliminarily results, considered as starting points for the validation of the pre-pilots. These results showed promising outcomes. It is expected that experiments with larger data sets will provide more insights of the methodologies and tools to be developed in the next cycle of the project.

In order to conduct the pre-pilots, currently, we are working on the knowledge elicitation process. A description of how the required data sets and samples from the medical domain have been generated. A scheme for annotation and grading of online discussions has been decided, and a think aloud protocol has been agreed, and sessions toward learner’s knowledge elicitation are carrying out at Manchester University Medical School.

Our initial steps towards the validation of the pre-pilots are promising and we expect that during the next cycle of the project we will build up from the collective experience accumulated during this first cycle.
5 References


D 4.1 Positioning Design


6 Appendixes

6.1 Appendix – 1: Description of the Pre-pilots

To exemplify the kind of work that will be performed within task 4.1 and task 4.2, this Appendix presents the showcases that will be elaborated. For a description of an initial set of experiments and test results see (Appendix-2 and Appendix-3).

6.1.1 Task 4.1. LSA Based Learner Positioning in Relation to his ePortfolio and Learning Materials

This show case presents a set of experiments aimed to evaluate the performance of LSA for positioning and advising the learner on whether the text presented as evidence of prior knowledge is sufficient to skip the study of a learning material.

Language: German
Domain: Computer science/IT
Nr of students: 5
Nr. of staff: 1

<table>
<thead>
<tr>
<th>Pre-pilot scenario</th>
<th>Long life learners want to achieve certain learning goals when they consider enrolling in curricula offered by an education provider. They have different educational backgrounds that they want to be taken into consideration.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The learner’s educational background can be of a formal nature (certified exams) in which case standard admission / exemption procedures may apply. In other cases such standard procedures are not available so assessors need to evaluate themselves the learner knowledge on specific topics.</td>
</tr>
<tr>
<td></td>
<td>In procedures for Accreditation of Prior Learning (APL) assessors decide whether evidence brought forward by the student may lead to exemptions from one or more courses.</td>
</tr>
<tr>
<td></td>
<td>A ePortfolio is a collection of documents (in the broadest sense) that may serve as evidence that the learner has mastered certain skills / can operate on a certain level of competence. The concept is based on the ePortfolio of the photographer / the model / the designer. Here we can think of the ePortfolio holding items such as: courses completed; readings by the learner; products of the learners (here mostly written stuff).</td>
</tr>
<tr>
<td></td>
<td>This pre-pilot validates a specific Latent Semantic Analysis (LSA) functionality (within a wider framework of a ePortfolio analysis service) that automatically analyses a learner’s ePortfolio for (1) advising the learner about the relevancy of material in the ePortfolio for potential exemptions in the curriculum; (2) analyzing knowledge gaps of the learner</td>
</tr>
</tbody>
</table>
There is an important assumption underlying this pre-pilot related scenario: similarity in learning materials can be used as a proxy for similarity in learning outcomes.

To validate the service functionality we will compare the service automatic recommendations for each text presented as evidence with expert recommendations over the same text.

**The objectives**

The objective of this pre-pilot is to validate the LSA based service functionality for assessing the learner’s prior knowledge on specific computer science related subjects and to advise the learner if the evidence within his or her ePortfolio is sufficient to skip the study of a set of learning materials associated to a curriculum.

The assessment is done by comparing a set of learning materials (gold standard) to a set of documents from the learner's ePortfolio that he or she provides as evidence of that prior knowledge.

**The Narrative**

The learner intends to enrol in a curriculum (here a series of learning objects) and then want to apply for exemptions based on APL.

Before the learner decides to enrol he or she wants to know if documents within his or her ePortfolio are sufficient evidence as for assessor to grant him an exemption on the relevant learning object.

Therefore the learner submits documents from his ePortfolio to the LSA based analysis service functionality to know if that document can be used as evidence for skipping a specific learning object.

The learner input to the service a) a document from his/her own ePortfolio; b) learning materials associated to the learning object.

Each evidence document and each learning material will consist in texts of about one paragraph long.

The functionality outputs are automatic recommendations based on a comparison of similarity between each document presented as evidence by the learner and the learning materials. Recommendation for “skipping” implies that the learner presented evidence was sufficient. Recommendation for “studying” implies that the learner presented evidence was insufficient.

**Interactions(s)**

An expert will configure the LSA related tools that will (a) generate a term to document matrix based on a domain specific text collection, (b)
decompose that matrix using the Singular Value Decomposition algorithm.
(c) reduce the dimensionality of the decomposed matrices generating the
LSA semantic space.

The same expert will train the LSA based functionality service by first
folding in the semantic space a set of pairs that includes a learning material
text and ePortfolio document text that has been pre-selected as good
evidence of prior knowledge for the topic covered by the learning material;
and second measure the calculating the average cosine similarity between
the texts within the pairs.

The aim of the training is to find an average similarity value between
document and learning material pairs that can be used by the service as
similarity threshold when the learner presents new pairs for evaluation.
Therefore the service will recommend that the document is a good evidence
for a set of learning materials when the similarity between the document
and each of the materials are higher or equal to the value of the threshold.

The learner will input to the service ePortfolio a pair of document and
learning material.

Then the service will automatically recommend that the document is
sufficient or insufficient evidence for that learning material.

An assessor will evaluate the document and will confirm if the automatic
recommendation is correct. The assessor will annotate the evidence
document within the pair based on his decision (i.e. relevant or not relevant
for learning material x). The annotated document is archived.

The next time a learner input a pair the service will measure the similarity
between the evidence document and the learning material within the pair
and, in addition, between the evidence document within the pair and the
annotated and achieved documents relevant for the learning material within
the pair.

The service automatic recommendation for each pair will be calculated
based on the observation that documents to be considered good evidence
should be more similar to archived documents annotated as relevant for the
learning material that to archived documents annotated as irrelevant for
the same learning material. Moreover the evidence document within the pair
need to have a similarity value high than the threshold obtained during the
training period.

<table>
<thead>
<tr>
<th>Used technology and services</th>
<th>R-package</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing opportunities and requirements</td>
<td>Context</td>
</tr>
<tr>
<td></td>
<td>Domain: Computer science</td>
</tr>
<tr>
<td></td>
<td>Language: German</td>
</tr>
</tbody>
</table>
Resources
A collection including 94 texts (essays) written in German by students in the about computer interfaces. The texts have been graded by tutors/experts using a scale from 0 to 4. The majority of those texts are approximately one paragraph in length. Some of them are shorter. In addition to the graded texts, each collection includes a set of 3 gold standard texts.

Corpora to build the semantic space:
2444 documents where 2/3 of the documents are generic from the press. 1/3 of the documents are domain-specific (i.e. textbook)
Size: 444k words
Document length: mean = 181, sd = 156
Term frequencies: mean = 7.4, sd = 120

Participants
1 expert

Plan for testing
From an available collection of student essays the expert selects two set of texts, a first set includes texts with low grades (<2) and a second set includes texts with high grades (=>2). Then each set is divided in a training set and a test set. Those subsets include an equal number of randomly selected texts.

The training subsets will be used to find an optimal set of parameters values (number of LSA dimensions, number of stems used to build the semantic space, similarity threshold value for the comparison between evidence document and learning material).

A domain corpus is used to build a semantic space. Both training subsets of the low grades and high grades are folded in to the semantic space. Then we measure the similarity between:
   a) each text (belonging to the first and the second subset of the test set) and its corresponding set of gold standards.
   b) each text and an increasing number of the rest of the texts. For each iteration we can compare the same texts with more texts and then obtain an average. This will tell us if the service can learn from the texts previously accepted and rejected by assessors (we use the high and low grades as a proxy for accepting or rejecting document as evidence).

Then, the testing set will be used to cross validate the best performing parameters obtained with the training set. The evaluation consists of measuring accuracy and precision.
6.1.2 Task 4.2. Monitoring Learner’s Conceptual Development in the Medical Domain

Describing a case in a chosen topic (and pre-specified format) and getting a LSA-based / concept mapping tool based feedback-report allowing self assessment of strengths and weaknesses in the concerning topic.

Language: English  
Domain: Medicine  
Nr of students: 5 (max 10)  
Nr. of staff: 1 (or 2)

<table>
<thead>
<tr>
<th>Pre-pilot scenario</th>
<th>4.2. Monitoring learner’s conceptual development in the medical domain</th>
</tr>
</thead>
</table>
| The objectives     | The main objective of this pre-pilot test is to demonstrate the application of language technologies to analyse a learner’s conceptual development. More precisely it will:  
  • Show until what extent language technologies can be used to monitoring the conceptual development of the learners  
  • Compare and set aside metrics derived from concept maps based tools with direct diagnosis of conceptual development of textual output  
  • Evaluate what can already be used with the current stage of available language technology based tools and what should be developed  

By directly using language technologies to analyse the learner’s conceptual development, our purpose is to –either– reduce the constraints imposed by concept-mapping based tools and therewith extend the number of situations in which learners can obtain feedback on their conceptual development –or– enrich the feedback possible with concept-mapping based tools.  
The main actors are:  
- learners  
- tutors |

| The Narrative | 1. Medical students will be asked to describe a selected medical case (diagnosis, symptoms, treatments, etc.). (Note: existing texts may be used, in which case the students are only involved indirectly). One of the following inputs will be used:  
  • Think aloud protocols provided by UMAN-students and converted into written text  
  • Available think aloud protocols (English and Dutch) from (van de Wiel, 1996) |

LTFLL -2008-212578 58
2. A (UMAN-)tutor will provide a set of typical case descriptions for the selected case. It should represent the level of knowledge a student should have at this stage.

3. The resulting text of students and tutor will be (manually) converted to fit into a concept map based analysis tool and subsequently analysed with this tool.

4. The resulting text of students and tutor will be analysed and compared with the help of INFOMAP/LSA with an initial set of indicators, e.g. co-occurrence of concepts, relations between concepts, missing concepts etcetera (c.f. deliverable 4.1).

5. The WP4-members will compare and analysis the results of step 3 and 4 and compile a report for the tutor.

6. The tutor will give feedback on the usefulness and the quality of the report.

**Interactions(s)**

1. Student receives a case and instructions on how to respond to this.

2. Tutor receives a case descriptions and instructions on how to respond to this.

3. WP4-members analyse the case descriptions using a) use concept map based analysis tools (i.e., HIMATT) and b) INFOMAP/LSA. After, they compare results from a) and b) and create a report for the tutor.

4. Tutor receives a report and instructions on how to respond to this.

**Used technology and services**

- INFOMAP or LSA (or maybe in combination with keyword extraction of LTs4eL)
- Concept mapping tools: HIMATT (SMD, MITOCAR) -or Pathfinder Analysis and KNOT.
- Corpora: general medical corpora (e.g., MedLine), materials of the topic (i.e., tutor notes, learning materials, keywords, text version of the use case, etc.), model answers.

**Testing opportunities and requirements**

Test environment:

- HIMATT (SMD, MITOCAR), or Pathfinder Analysis and KNOT, INFOMAP/LSA
- domain: Medicine
- language: English

Resources (see previous section)

- Two computers

Where the testing will be done: UMAN-OUNL with UMAN; WUW IPP-BAS (depending of language and input)

Plan of action (for testing the pre-pilot only):

1. Elaboration of a structured interview to be used in the think aloud
sessions. The objective of the interview is to obtain the data needed to evaluate the conceptual development of the learner.

2. Perform a pre-test of the tools that will be used in the pre-pilot concept map based analysis (i.e., HIMATT) and for the extraction of conceptual maps, INFOMAP/LSA.

3. Define how the pre-pilot will be run in terms of:
   - The way the tools will be used, or combined
   - The different methods to calculate the similarity of concept maps (e.g., a concept-based approach, a link-based approach)
   - The set of indicators that will be used to analyse and compare the results, e.g., co-occurrence of concepts, relations between concepts, missing concepts, etc.

4. Find out the corresponding corpus of learners’ output texts and tutors’ texts; conducting the think aloud sessions and convert the output into text.

5. Convert the texts so they fit into the HIMATT tools and analyse the results according to a set of indicators.

6. Analyse and compare the text using INFOMAP/LSA and analyse the results according to a set of indicators.

7. Compare the results of step 3 and 4 and generate a report.

8. Ask tutor to give feedback on the usefulness and the quality of the report.

### References

| HIMATT tools | http://himatt.ezw.uni-freiburg.de/cgi-bin/hrun/himatt.pl |
| INFOMAP      | http://infomap-nlp.sourceforge.net/doc/contact.html |
| KNOT         | http://interlinkinc.net/KNOT.html |
6.2 Appendix – 2: First Experimental Results Task 4.1

Positioning requires mapping learner knowledge onto a set of required competences associated to specified learning goals. In this preliminary experiment we evaluated performance of LSA and more formal approaches (e.g. semantic search). We used 96 essays and 3 gold standard (also essays) representing the learning goals of students for a particular topic in Computer Science (i.e. interfaces). The 96 essays were partitioned into accepted and rejected. We compared machine judgments against human judgments and used LSA to calculate these machine judgments (see Figure 6).

![Graphs showing frequency distribution and normal Q-Q plot for document length and density.](image)

Figure 6
Facts about used corpus:
2444 documents
  2/3 generic: press
  1/3 domain-specific: textbook
Size: 444k words
Document length: mean = 181, sd = 156
Term frequencies: mean = 7.4, sd = 120

We used human scores and machine scores. S = 67528.3, p-value = 1.319e-07. Alternative hypothesis: true rho is not equal to 0. The sample estimates have a rho = 0.5121316. The medium effect is highly significant but still two humans graders (untrained) are also often around 0.6 (see Figure 7).

Figure 7. Recall and precision using different rules and different mix of accepted and rejected essays
The experiments were run with five different settings each containing a randomly chosen training and test set. For each experiment, precision and recall were measured to find out if an algorithm can learn from previous inputs and if it is better or worse compared to the others. For that reason, five different learning rules of an algorithm were defined for testing if an essay is good or bad evidence.

- **Rule 1 (Best of three golden):** The threshold is computed by averaging the similarity of all three golden standard essays to each other. The similarity of the investigated essay is compared to the best three golden standard essays (=machine score). If the machine score correlates above the threshold with the human judgment, the test essay is stated correct. This rule assumes that the gold standards have some variation in the correlation among each other and using the average correlation among the gold standards is taking that into account.

- **Rule 2 (Best of goods):** Best essays of the humanly judged good ones. The assumption behind this is that the similarity between these good examples is getting more precise with rising numbers of positive evidences. With more positive examples to evaluate an investigated essay against, the precision of the evaluation should rise. The threshold is the average of the positive evidence essays among each other.

- **Rule 3 (Average to good > average among good):** Test if the similarity to the good ones is higher than the average similarity of the humanly judged good ones. Assumption is that the good evidence gathered circumscribes that area in the latent semantic space which is the model solution and that any new essay should be within the boundaries characterized by this positive evidence thus having a higher correlation to the positive examples than they have among each other.

- **Rule 4 (Best of good > best of bad):** If the maximum similarity to the good essays is higher than the maximum similarity to bad essays. If a tested essay correlates higher to the best of the good than to the best of the bad, then it is added to the class of good ones.

- **Rule 5 (Average of good > average of bad):** Tests the same as average of good > average of bad. Assumption behind this is again that both bad and good evidences circumscribe an area and that the incoming essay is in either the one or the other class.

Results show that:

1. Recall and precision stay stable as no changes to the reference material are done: all essays are evaluated the same way as the training material is not taken into account.
2. Recall goes down when having less good essays in the training sample; recall is much higher than in best of gold; precision stays the same; with enough good examples, recall is higher than in rule 1.
3. Recall with 80% good is same as with 20% good, very volatile (also compared to two additional randomizations in the outer loop); precision very stable and a bit
higher than in the previous rules; it seems to be very dependent on the type of positive examples whether they are able to characterize representative boundaries: seeing recall change with varying amounts of positive examples, this indicates that the boundaries are not very well chosen.

4. Recall goes down significantly (but starting at a very high level), precision is relatively stable; having more negative evidence is counterproductive and it is more important to have positive examples; interpretation: first, bad essays scatter across the space and it is likely for a good essay to correlate higher with a bad one when there is only a low number of positive evidence; second, bad essays might contain very few words and thus expose correlation artifacts (easy to detect, but not with LSA).

5. Recall is generically higher than in the best of gold cases; precision is in the same area; recall seems not to be so stable but does not drop with more bad samples such as in the best of good case; interpretation: noise can be added to increase recall while still only a low number of positive examples is available.

Conclusion

Distractors are of low value in the rules tested. It seems that generic noise can be added to keep recall higher while only a low number of positive examples can be utilized. An explanation for this can be found therein that there are always a lot more heterogeneous ways to make s.th. wrong so homogeneity can only be assumed for the positive evidence, not for negative evidence.

Noise seems to be useful though for the calculation for the threshold. Though it will need further investigation to confirm our hypothesis, we believe that learning will help. Recall could be improved in various cases, while precision stayed at the more or less same level. Thresholds and ways how to calculate them are evidently important. We proposed several well working ways on how to construct thresholds from evidence that extend the state of the art. Thresholds usually vary with changing corpus sizes and the measures proposed can adapt to that.

An open problem remains the small sample size for train and test groups, which can lead to errors. No new findings about the critique that some concepts may not be realized with words (but e.g. images).

We were able to show that textual representations of free-text with few words works. There is related work that indicates LSA can be applied for smaller textual units (Landauer’s Toefl test experiments) as well as longer textual items. A meta-study on evidence length would be useful.

By using a seed set of three model solutions, we were able to show that by using a textbook split into paragraph sized textual units combined with generic background material, valid classifiers can be built with relative ease. Furthermore, reference material to score against can be collected along the way.
6.3 Appendix – 3: First Experimental Results Task 4.2

Based on a psychology dataset from the OUNL several small experiments have been done. The dataset offers course content\(^\text{11}\) (=expert level argumentation) and learner’s documents. For every part of the course content we know some related keywords based on the authors information.

1. Keyword relations

Since we know the keywords for the psychology chapters, the inter-correlations of the keywords have been calculated for one exemplary chapter.

2. Clusters in these keywords

With a clustering method (nearest neighbour approach) we have produced a distance matrix and clusters in these keywords. Here you see an exemplary cluster overview for the same chapter as used in 1 (see Figure 3).

3. Semantic Networks around these keywords

For every keyword we have calculated a semantic network based on Infomap. All keywords are used as query terms so an extended semantic/concept map around these keywords is built.

4. Semantic networks from student documents

The author provided keywords have been taken as query terms to build semantic networks from the student documents.

First results: Since the student texts have not been provided from a study context but more from several contexts the dataset does not seem to be ideal for such a kind of analysis. The level 1 keywords are not found too often in the student documents, but it would be still an option to go for a proxy-by-proxy approach and take the highly related keywords from the expert documents.

\(^{11}\) Notice that in this appendix “expert knowledge” has been defined as the course content. The “expert knowledge” can be also seen as the knowledge a learner, who is considered by the tutor as an expert learner, has in a particular context.
6.4 Appendix – 4: Test of the ALA-reader, CMAP and KNOT Tools

This appendix shows a testing of the of the Ala-reader software and KNOT (Clariana et al., 2006), and CMAP (Institute for Human and Machine Cognition, 2004).

**Procedure**

1. Download Ala-reader.exe and it files from [http://www.personal.psu.edu/rbc4/score.htm](http://www.personal.psu.edu/rbc4/score.htm)
2. Check whether all files are correct (sometimes the newlines are wiped out by downloading) this can be done by inspecting the terms.txt.
3. If files are incorrect go to the files at [http://www.personal.psu.edu/rbc4/](http://www.personal.psu.edu/rbc4/) and open these in the browser and use the copy and paste functions to ensure your local files are correct.
4. Open alareader.exe and use as first try the file ‘expert1.txt’ if this gets a maximal score continue with the essay files if not go back to 2. and correct the files
5. When you change the terms in the term file ensure that each line has two forms (normally one abbreviation or synonym), otherwise the terms are not recognized.
6. The L_report.txt file will be extended with the results found from each new evaluation.

The ala-reader is a blackbox, which is developed with Authorware. The behaviour is partly described in the papers of Clariana et al. (2006) and partly deduced from small experiments using the tool itself.

How to get the number of terms? It was assumed that the literal terms from the terms.txt file are used to just count all found ones.

How to get the number of propositions? Count the terms in each sentence. Each combination in a sentence (or on one line) is a proposition. Each proposition receives a specific label. For the essay all propositions are stated, the combinations which are identical with the expert ones, they get the same label. Afterwards, it has been counted how many propositions are shared with expert1 and expert2.

**REMARKS:**

1. No syntactical information is used, two separate words are processed on the same way as a sentence with both words in it. (“lungs oxygen clean” or “The lungs and the oxygen combined with clean”). Also the word order and the possible relations will be neglected “Lungs contain oxygen” is the same as “Oxygen contains lungs”
2. If all terms are put in one line, has one completed match with the expert? No, because not all combinations are made (the process terminates somewhere, it loops only once through the line). The Ala-reader gives 1 agreement back with the expert, which is incorrect (should be 2). If this is done manually, making propositions between term1 and term2, term2 and term3, etc. and a comparison of these is performed with the expert, two correct matches are found (“aorta-a1-aortic valve” and “aorta-a2-body”). The first one is missing in the output file. This is not because the position in the essay file (new first word does not solve the problem). The hypothesis, then, is that within the search in the expert propositions (to fill the proposition labels) the double loop is not correct.
3. The counting of the propositions is not correct. The “cat and dog in the car” gives 2 instead of 3 propositions and the Ala-reader even reduces the match to 1 (see previous remark).
4. It does not use lines for the counting. “The lungs need the oxygen it contains clean.” returns 2 propositions if it is written as one line or if it is spread over 8 lines.
5. The expert1.txt was used as an input and it gave back exactly 29 (as the Ala-reader tells) propositions as output in a sorted list.
6. The expert2.txt was used as input and it gave back exactly 32 (as the Ala-reader tells) propositions as output in a sorted list.

7. It seems that the memory is not complete cleaned between different runs with the same file because it gives different output files. Not by extending the file (after checking with a half empty file) but from the memory in Authorware. Three times evaluating the creation of the output file that all found propositions are listed three times (in the L Cmap data folder) \Rightarrow close Ala-reader between each new evaluation.

8. The memory is also not cleaned with two different files, all propositions of the previous one are also written into the output file (in the L Cmap data folder) \Rightarrow close Ala-reader between each new evaluation.

After some experiences with the Ala-reader the output has been imported into CMAP with the following results:

<table>
<thead>
<tr>
<th>CMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert1</td>
</tr>
<tr>
<td>Essay01 (1-1)</td>
</tr>
<tr>
<td>Essay04 (4-4)</td>
</tr>
</tbody>
</table>

LTfLL -2008-212578

67
Remarks on CMAP:
1. A visual comparison of these CMAPS are quite cumbersome, see the examples in the table above.
2. The same input files gives not the same concept map layout see below for 4 different tries:

3. The data files created with the Ala-reader are incorrect to be used for the KNOT software. The 26 terms should create a matrix with 25 lines of zeros and ones. However all files have 29 lines. So first after that the last four lines were deleted KNOT produced the following maps:
Here you see that visual inspection produces the same problems as CMAP does. However, the KNOT software gives the functionality to compare all the networks in one run. This was performed, and the KNOT software produced the following tables:
Comparing all the networks to each other, it has been found that:
Remarks
1. Looking to the produced tables and matrices, it is supposed that the sim and csim matrices provide the best indication of the quality.
2. KNOT collapses if concept maps are combined in order to make a visual comparison.
6.5 Appendix – 5: Review of Systems Performing Sentiment Analysis

**Jane16**
Description: Java Associative Nervous Engine is online content text analyser, which include downloadable open source sentiment analysis program. Jane16 implements number of algorithms like Bayesian model, K-means clusters, many sort of filtering algorithms etc. in one single system that is capable to extract online the relevant information of the text like subject, summary and in certain cases even sentiment of the text.

**LingPipe**
Site: [http://alias-i.com/lingpipe/index.html](http://alias-i.com/lingpipe/index.html)
Description: LingPipe is a suite of Java libraries for linguistic analysis. LingPipe's information extraction and data mining tools perform the following tasks: track mentions of entities (e.g. people or proteins); link entity mentions to database entries; uncover relations between entities and actions; classify text passages by language, character encoding, genre, topic, or sentiment; correct spelling with respect to a text collection; cluster documents by implicit topic and discover significant trends over time; and provide part-of-speech tagging and phrase chunking.

**OpinionFinder**
Site: [http://www.cs.pitt.edu/mpqa/opinionfinderrelease](http://www.cs.pitt.edu/mpqa/opinionfinderrelease)
Description: OpinionFinder is a system that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. OpinionFinder operates as a pipeline. First, some standard text processing operations are performed (e.g., tokenization, POS tagging). They are followed by a subjectivity analysis: subjective sentence classification, direct subjective expression and speech event identification, opinion source identification and finally, sentiment expression classification. The output files are in the form of SGML/XML markup of the original documents. This system is available only for Linux.

**RapidMiner**
Site: [http://rapid-i.com/content/blogcategory/38/69](http://rapid-i.com/content/blogcategory/38/69)
Description: RapidMiner (former YALE) is an open-source Java software that covers a wide range of data mining tasks.

Sentiment “dictionaries” (tagged word listings) for English:
Inquirer Dictionaries
Site: http://www.wjh.harvard.edu/~inquirer/homecat.htm
Download: http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
Description: The Inquirer Dictionaries include words grouped in different tag categories depending on the source: the Harvard IV-4 dictionary, the Lasswell value dictionary, and "marker" categories, presented in Kelly and Stone (1975). In addition, there are two valence categories, tagged Positiv (1915 words) and Negativ (2291). The remaining 7582 terms that are tagged neither as Positiv, nor as Negativ, can be considered neutral or objective.

Subjectivity Lexicon
Site: http://www.cs.pitt.edu/mpqa
Download: http://www.cs.pitt.edu/mpqa/hltemnlp05clues.tar
Description: The Subjectivity Lexicon is part of the OpinionFinder system. This dictionary contains 8221 subjectivity clues both manually and automatically identified (Riloff and Wiebe 2003). Each of the clues is a single word, marked among other features for its out-of-context prior polarity (positive, negative, both, neutral) and type (strongsubj, when the word is subjective in most contexts, otherwise – weaksubj).

Micro-WNOp corpus
Site: http://www-1.unipv.it/wnop
Download: http://www.unipv.it/wnop/micrownop.tgz
Description: The Micro-WNOp corpus is composed by 1105 manually evaluated WordNet synsets (Cerini et al. 2007)
6.6 Appendix – 6: Text Available for Analysis

This appendix describes learners’ text that will be available for testing the tools and services that will be developed.

The learner population is medical students of the University of Manchester. For task 4.1, positioning the learner, text material generated in the curriculum in online discussion will be used, whereas for task 4.2, conceptual development, and a think aloud protocol will be followed.

Background of the curriculum

The present integrated curriculum in the University of Manchester Medical School is founded on problem-based learning (PBL), which encourages self-directed study. This approach was introduced in 1994. Students now receive a four-week induction in Essential Skills at the beginning of the programme to prepare them for PBL. The five-year course is made up of three phases and an assessment activity performed over the course of the programme (see Table 7).

<table>
<thead>
<tr>
<th>Phase</th>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, 2</td>
<td>Phase 1 is specifically designed to prepare students for problem-based learning and to lay a firm foundation of knowledge, skills and attitudes in the biomedical, clinical, behavioural and social sciences underlying medicine. This knowledge and understanding is contextualised through interactions with patients, communication and clinical skills training throughout years 1 and 2. At the end of each module, students undertake a Student Selected Component (SSC), which allows students to pursue a particular area of interest, undertaking a placement and writing a report about a patient or condition within a chosen area. Students begin to compile a portfolio from the start of Phase 1, as described in the previous section.</td>
</tr>
<tr>
<td>2</td>
<td>3, 4</td>
<td>In Phase 2 students revisit the topics studied in Phase 1 within a clinical workplace learning environment. Students continue with SSC’s as in Phase 1. Phase 2 ends with the Project Option, where students study a particular problem, speciality or aspect of medicine in depth. Students choose between undertaking research or a detailed clinical audit within a particular area, culminating in a written report and a presentation.</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>Phase 3 is the transition period for students to become doctors. There is a strong emphasis on clinical apprenticeship and learning almost entirely from live patients in real clinical situations. Phase 3 students are required to undertake five modules: Elective, Community, Teaching hospital, District general hospital, Consolidation.</td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td>Students undertake a range of summative assessments over the course of the programme, which include practical assessments of clinical skills. Assessments of knowledge and understanding follow the general format of solving a clinical problem by selecting the appropriate response from a menu of possibilities. The student responses are optically read and the results are recorded electronically.</td>
</tr>
</tbody>
</table>

Table 7. Curriculum structure of University of Manchester Medical School
Text material generated in the curriculum

Student Selected Components (SSC): SSC’s components culminate in a 3,500 – 4,500 word structured report, demonstrating significant exploration of a specific area of medical knowledge. The student explores their own particular area of interest. Each SSC topic has a specific tutor who completes a feedback and assessment form for the student. These are now being introduced in an electronic format. For comparative analysis using LSA, this would result in limited correlation as each student will study a discreet set of concepts. Each report contains a set of references, which could be harvested and used in corpus-based analysis, however. Furthermore, as students complete SSC’s throughout all 3 phases of the curriculum, it might be possible to use analysis of language to investigate a particular student’s learning development.

Online discussions: Asynchronous online discussions were introduced into Phase 2, to continue the discussion of issues concerned with personal and professional development, which were initiated in face to face groups in Phase 1. Increasing student numbers and their geographic dispersal required the introduction of electronically based discussion fora. Students work in facilitated groups to discuss specific contemporary issues throughout the curriculum, with a nominated student taking on a facilitator role. The asynchronous nature of the discussions enables the participants to respond reflectively and critically in the context of their own experiences. There are 63 discussion groups in each of years 3 and 4, consisting of 8 students, including the student facilitator. Each text flow usually consists of at least 8 – 10 postings. Although some of these are of just one or two lines, mostly they consist of at least 10 lines of text. All discussion groups participate and each discussion topic stimulates over 2,000 postings.

These discussion fora provide ideal datasets for LSA, both for comparative analysis and for corpus based analysis. For learner positioning, this provides some interesting possibilities across the groups, to suggest concepts that other student groups may be exploring, for example. It may also be possible to determine the progress of individuals within a particular group. An example is provided by discussions on the safe prescribing of medicines. This topic was initiated by suggestions to the students of the type of knowledge based material on which to base their discussions, including some detailed compendia of drugs and treatment e.g. the British National Formulary. This may provide a sound corpus on which student discussions can be analysed and feedback provided to ensure good coverage of the subject area.

Data generated by “think aloud” analysis of PBL cases

There is a major need to provide structured formative feedback to the students as individuals of their conceptual development, resulting from their learning in their PBL sessions. An appropriate means of achieving this is to record a student’s brief analysis of a recent PBL case and its relevance to the clinical condition concerned. The transcribed text output is suitable for language analysis that will allow a grading on which feedback to the student can be based. The feedback will focus on their understanding of the subject area and their ability to relate it to the clinical problem originally posed. This method of text generation is especially appropriate to the medical domain as clinicians routinely provide verbal reports of clinical problems to colleagues, which often include a range of healthcare professionals.