Shifting Interests: Changes in the Lexical Semantics of ED-MEDIA

Fridolin Wild, Chris Valentine, Peter Scott
Knowledge Media Institute, The Open University, UK
{f.wild, c.p.valentine, peter.scott}@open.ac.uk

Abstract: Large research networks naturally form complex communities with overlapping but not identical expertise. To map the distribution of professional competence in field of 'technology-enhanced learning', the lexical semantics expressed in research articles published in a representative, large-scale conference (ED-MEDIA) can be investigated and changes in the topics covered can be tracked over time. Within this contribution, the two years 2000 and 2008 are used to contrast the change in meaning structure. In both cases, first a quantitative investigation is applied, directly followed by a two-fold qualitative analysis. Results indicate that the field has opened up to new areas, thus diversifying research; at the same time, certain terminology disappears as interests shift and research is finalised for some of the underlying problems.

1 Introduction

Research is a process largely relying on self-organization. In a way, communicating research itself is a negotiation process aimed at reaching an agreement about knowledge and about who has this knowledge. Large research networks therefore naturally form complex communities with overlapping groups in varying sizes. It is clear that an individual's efforts within this network cannot be independent of the whole. Is it not least through influence and identification that this network interacts.

Both professional and rich professional competences of the actors are necessary for such a research network to be successful (cf. Rychen & Salganik, 2003). Professional competences relate to the expertise of the network: they refer to the potential to construct domain-specific knowledge within the network or within parts of it. Rich professional competences transcend these domains and encompass – amongst others – social competence, self competence, and methodological competence. Social competence, for example, refers to the potential to undertake (collaborative) actions in order to identify, manage, and master conflicts (see Erpenbeck & Rosenstiel, 2003).

Evidence for inspecting the characteristics and for estimating the distribution of professional competence is reflected in the various research outputs produced in the scientific value creation chain: peer-reviewed media such as conference, journal, and workshop publications, in the recent years additionally also (non-peer reviewed) online articles and blog postings. Prizewinning and keynote activities provide more strongly weighted data sources. Funding data and information about joint projects can complement the picture.

Evidence about the distribution of rich professional competences on the other hand cannot be derived directly from the artefacts produced. It has to be inferred indirectly from the actions and relationships of the protagonists of the network: most notably their collaboration in authoring, but also their joint attendance in events and meetings, or their affiliation in organisations and special interest groups can serve this purpose.

This analytical work is focussing on creating higher professional awareness about the shifts the field technology-enhanced learning as such is experiencing over time: new topics emerge, old research strands finish, major topics become minors – and minors majors.

For this contribution, we therefore have utilised qualitative and quantitative analysis methodologies to depict the changes in semantic structure of the field technology-enhanced learning – exposed in one representative, large-scale TEL conference: the ED-MEDIA. We have set aside people and organizational analyses for later work. Deliberately, we also neglected journal publishing for now, as – although having a much higher quality profile – the publication cycle is much slower. So whilst it may reflect a more mature consensus around field themes, these are harder to tie to specific points in time, as quality journals can have very heterogeneous and not rarely very slow publication cycles.

The rest of this contribution is organised as follows. A methodology section outlines the means used to conduct the subsequently following analysis. Therein, the two years 2000 and 2008 are contrasting the changes that took place in the field. In both cases, first a quantitative investigation is applied, directly followed by a two-fold qualitative analysis. In the next section, the findings are being discussed. The contribution is wrapped up by an outlook on possible extensions of this study.

Methodology

The terminology used to title research articles in general captures the essence of the contribution. When monitoring a larger number of contributions over time, changes in the use of this terminology can be detected that reflect shifting interests in the topics covered by a research network. These changes can be of very different nature.

Quantitatively, the growth of the dictionary over the years analysed is of interest. Overlaps in terminology, both newly introduced and disappearing keywords (aka ‘terms’) mark quantitatively, how the terminology shrinks or grows. Bursts in frequency facilitate the detection of shifting relevance of keywords, especially among the medium frequent ones that are considered most semantically discriminative.

Qualitatively, the structure of the semantic relationships in this dictionary is of interest: some terms are closer to others thus allowing to see the dictionary as a semantic network: nodes are keywords and links represent their weighted semantic closeness.

There are many competing models to automatically determine the 'closeness' of terms with the help of natural language processing. Among these models, latent-semantic analysis (Landauer et al., 1990; Wild, 2006) as an extension of the classical vector space model (Salton et al., 1975) has been shown to provide high performance. Mapping the terminology in a lower-dimensional, ‘latent-semantic’ vector space helps to measure the geometrical distance between terms as a proxy for their semantic closeness. For the two years 2000 and 2008, a comparison of these resulting graph structures has been conducted.

The publication data of the ED-MEDIA conference series has been partitioned by years. The titles of the contributions were sanitized by stripping all non alpha-numeric characters, converting all remaining words into lower case, and removing all words with less than two character length. A lower frequency bandwidth threshold was introduced, effectively eliminating all words appearing only twice or less in the titles. English stopwords (see Wild, 2008) have been removed on the other side of the frequency spectrum to keep pronouns and other functional terms such as 'the' or 'it' from distorting the analyses. The remaining vocabulary was aggregated along word stems using Porter's snowball stemmer (see Temple Lang, 2009).

The first step of the analysis focuses on the change in terminology. Therefore, the frequency of the words of this remaining vocabulary in the document titles is computed. Subsequently, the yearly change in vocabulary use is assessed: a tabulation of the normalised frequencies gives insight about the nature of the terminology (depicted by bar plots and log density curves). When comparing the tabulations of two years, the terms that disappeared, the terms that are new, and changes in distribution can be assessed. The changes in distribution can give insight in diminishing or enforced roles of keywords via a simple burst detection, i.e. a significant increase or decrease in usage frequency in the comparison data set.

Pseudo documents are created reflecting the frequency distribution of keywords in the classes 'new', 'gone', 'diminished', and 'enforced' are used as input for a wordle.net word cloud diagram that reflects higher frequencies with larger letter size.

In a second step, for each year a latent-semantic space is calculated by conducting and truncating the results of a singular value decomposition over a frequency table with the sanitised vocabulary used in this year in the rows and the document titles in the columns, thus having the frequency of each keyword in each document in the cells. As an estimator of a reasonable number of singular values to keep, dimcalc-share (Wild et al., 2005) is used. For each year, a distance matrix using cosine distances is calculated which serves as input to the divisive clustering algorithm Diana (Maechler, 2008). In the resulting cluster hierarchy, a reasonable cut-off point is estimated visually using the dendrogramm, and the tree is cut into a set of clusters.

For each cluster in each year, the graph component of this cluster is extracted from the graph, and a separate network plot (see Butts, Hunter, and Handcock, 2008) is created effectively linking the closest terms (using the all positive cosine distances as a weighted proxy). The 'backbone' structure of the component interactions is calculated by focusing on the single maxima in the directed incidence matrix of the cosine distances.

The deployed software is the language and environment R with the packages lsa (Wild, 2008), network (Butts, Hunter, and Handcock, 2008), sna (Butts, 2007), and cluster (Maechler, 2008). The analyses' R source code is available upon request from the authors.

3 Analysis

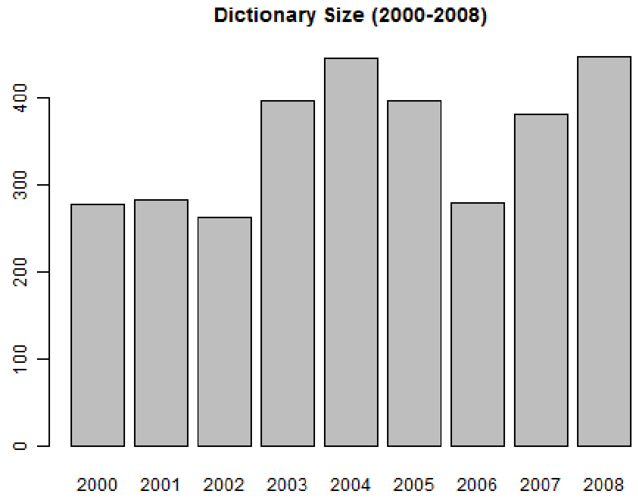


Fig. 1. Dictionary Sizes (2000 – 2008)

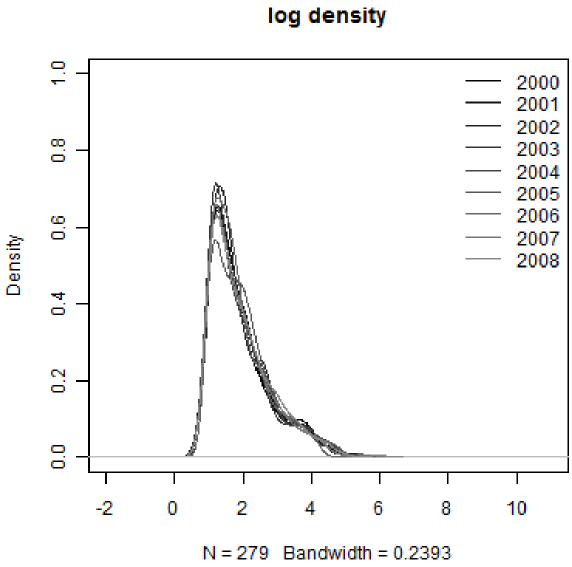


Fig. 2. Frequency Distribution (log density, 2000-2008)

Quantitatively, a generic rise in the dictionary size can be identified from the early years to the latter. Especially, the dictionary size of the year 2008 compared to the baseline 2000 has increased visibly (see Fig.1). After removing very rare terms and stopwords, 170 terms more had been used in the cleaned titles in 2008. Looking more closely, an overlap of 231 terms can be identified, whereas 218 new terms were found in the 2008 and 48 terms disappeared compared to the 2000 data-set.

Allover, the distribution of the frequencies stays relatively the same. Following Zipf's law, the medium frequent terms can be considered to be most discriminative regarding w.r.t. their meaning. In all years, there is a small number of very frequent terms (as stopwords had already been removed). Still, a larger group of low frequent terms is left behind by the rather tolerant lower-frequency threshold barrier of two (see Fig.2).

Analysing frequency shifts in the usage frequency of the terms in these dictionaries can provide insight in the terminological, lexical-semantic changes. To discover bursts, a just-notifiable-difference threshold of .001 was utilised. This threshold equals an absolute increase by 5.4 uses or a relative increase by 27.9 % of the mean frequency in 2000. Using this threshold, 14 terms seem to have a diminished role whereas 35 expose an enforced role.

A word cloud (using wordle.net) to visualise each of the four groups (diminished role, enforced role, new, disappeared) is depicted below, reflecting the actual frequency in the letter size.

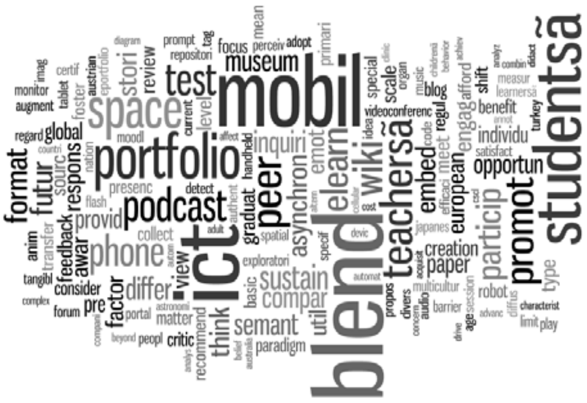


Fig. 3. New Terms (2000 to 2008).

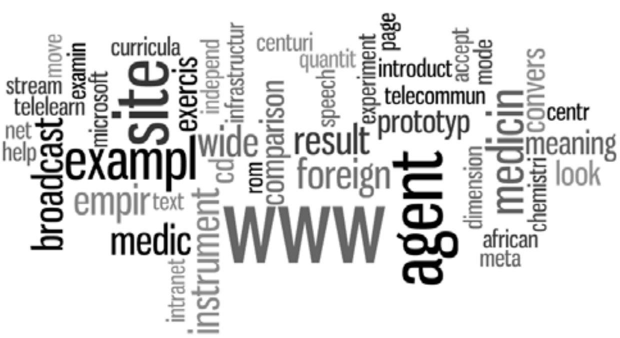


Fig. 4. Disappeared Terms (from 2000 to 2008).

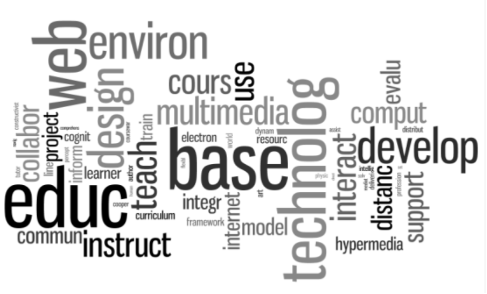


Fig. 5. Diminished Role (in 2008).

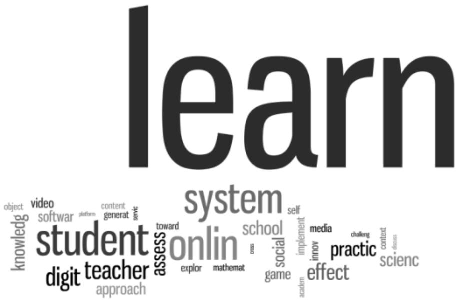


Fig. 6. Enfored Role (in 2008).

Qualitatively, through the cosine distances of the term vectors in the latent-semantic space, a graph structure can be established that – similar to a social network – connects terms as the nodes in this graph with each other through links, thus reflecting their semantic closeness. Using divisive clustering (precisely: the diana divisive analysis clustering algorithm), this network can be partitioned hierarchically into components. When looking at the dendrogrammes depicted below, a useful level of analysis for inspecting a reasonable number of components on the same level can be set at a height of .1. For the graph of the dictionary in year 2000, this effectively results in ten components, whereas 2008 lists 14 clusters.

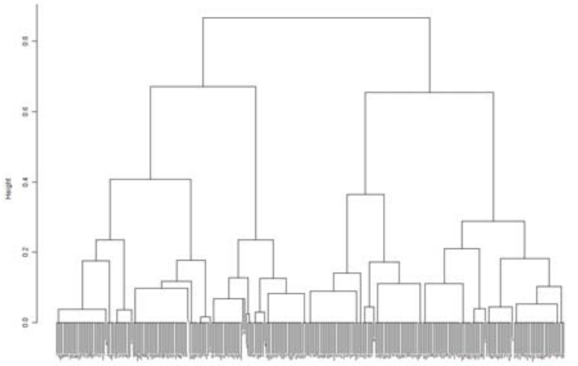


Fig. 7. Dendrogramme: Cluster hierarchy using divisive partitioning (2000)

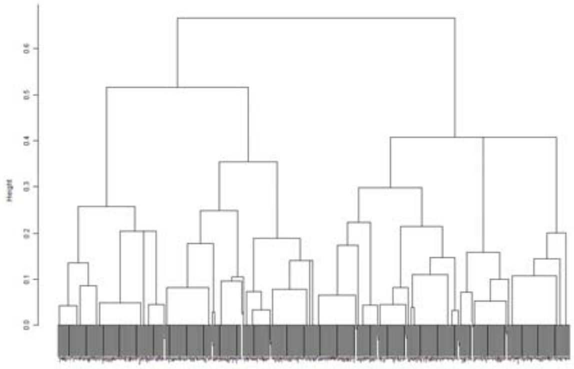


Fig. 8. Dendrogramme: Cluster hierarchy using divisive partitioning (2008)

Subsequently, the network structure of these components to each other and of the terminology networks within each cluster can be analysed in more detail (see Fig.s 9 and 10).

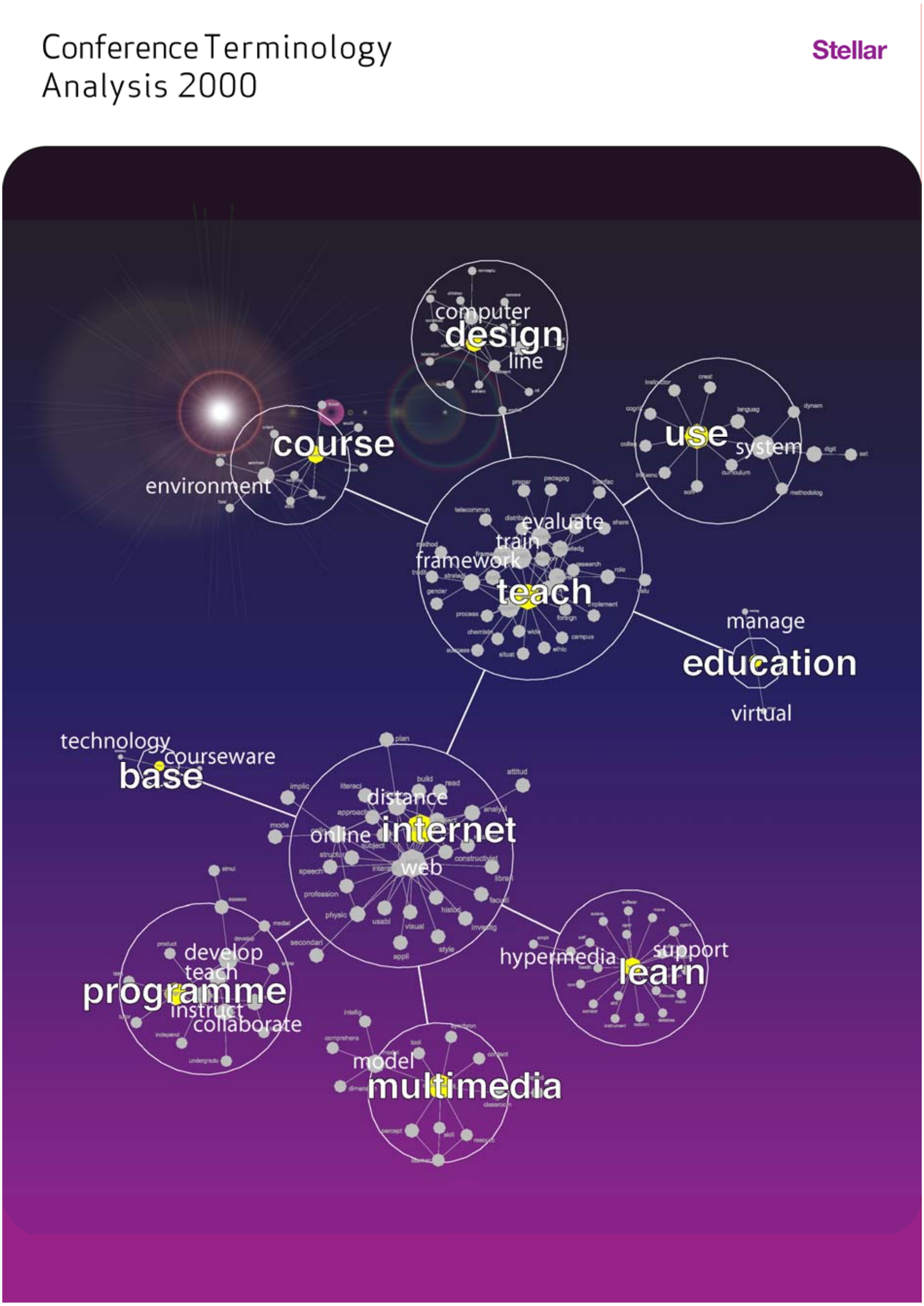
cluster	#	key terms
1	32	**use**, system
2	30	**learn**, support, hypermedia
3	27	**multimedia**, model
4	8	**base**, courseware, technology
5	23	**course**, environment
6	43	**teach**, framework, train, evaluate
7	30	**design**, computer, line
8	43	**internet**, online, distance, web
9	33	**programme**, instruct, teach, develop, collaborate
10	10	**education**, virtual, manage

Table 1. Components and their key descriptors in 2000.

cluster	#	key terms
1	35	**use**, skill, design, multimedia
2	39	**system**, distance, programme, media
3	48	**base**, structure, culture, campus, interact, (game), (wiki)
4	28	**technology**, community, approach, evaluate
5	22	**education**, social, effect
6	45	**student**, teach, peer, professional, service, (process), (mobile), (augment)
7	63	**environment**, tool, software, assess, manage, adapt, teacher

8	47	**web**, generate, value, success, ict, future, learner, change
9	39	**content**, **compute**, (virtual), (improve)
10	43	**practice**, study, collaborate, instruct
11	16	pedagogy, integration
12	15	**science**, **podcast**
13	5	**traditional**, **learn**, forum, read, flash
14	4	knowledge, train, strategy

Table 2. Components and their key descriptors in 2008.



Conference Terminology Analysis 2000

Stellar

Fig. 9. Lexical Semantics in 2000

Conference Terminology
Analysis 2008

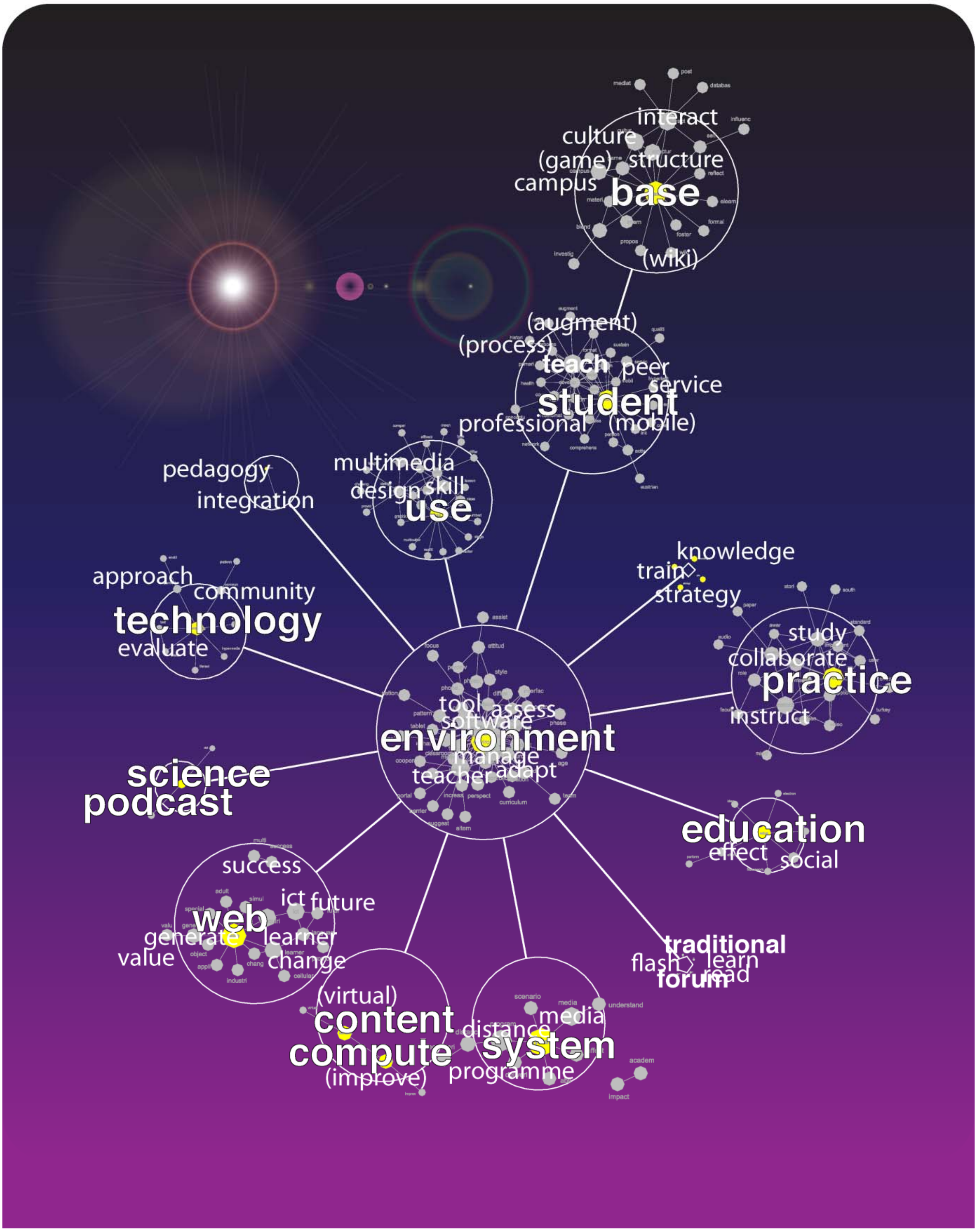


Fig. 10. Lexical Semantics in 2008

In 2000, the ten main components can be described with the terms exposing the highest betweenness (Butts, 2007; Freeman, 1979) in each component. Sorted by betweenness, these terms (2000) are listed in table 1. Similarly, table 2 lists the key descriptors for the year 2008.

Including the component networks, the year 2000 graph can be visualized as in the Fig. 9. For 2008, again including the component networks, the graph can be visualized as in Fig. 10.

4 Discussion

Comparing the two years 2000 and 2008, several interesting changes can be identified. First of all, – quantitatively –, there is an increase in dictionary size, though it should be mentioned that 79% of the terms re-appear in 2008 and only 21% of the terms disappear. The field seems to have broadened, the research results become more heterogeneous. It is not the case that this effect can be merely reduced to the bigger number of contributions accepted for publication (2008: 884, 2000:582), as the distributions have been cleaned from high- and low-frequent outliers and have been normalised in the further quantitative analysis process. This is also reflected in the component structure of the divisive cluster analysis: did a height of .1 in 2000 yield still 10 clusters, in 2008 already 14 clusters could be identified. In this time, the mean component size has increased in these years from 27.9 terms in 2000 to 32.0 terms in 2008.

Secondly, – qualitatively –, the change in the semantic structure can be characterised in the following way. As can be seen in the word cloud visualising the characteristics of the emerging terms from 2000 to 2008, the most prominent terms that appeared as new (in the order of their relevance expressed in their frequency) are: blended (28), ict (27), mobile (25), portfolio (16), space (16), peer (13), and podcast (12). Terms that seem to be clearly less important (diminished role) are: distance (-20), web (-18), hypermedia (-17), computer (15), internet (15), and multimedia (14).

At the opposite side of the spectrum, the following terms have become more frequent (numbers are extrapolated from the delta in normalised frequency counts), i.e. more important: digital (+33), teacher (31), practice (27), social (26), student (24), game (23), science (17), assess (15), effect (13), implement (12), innovative (12). The term 'learn' was excluded manually from this second list, as it is a too high frequent (+54, absolute 398 in 2008) term to be discriminative of meaningful changes. At the same time, the most prominent terms with a similar normalised frequency that disappeared are: www (13), medic(ine) (6+5), agent (9), and site (9).

In the network components generated by the divisive clustering over the term closeness, the semantic ‘backbone’ structure for 2000 can be circumscribed by focusing strongly on institutional needs and – viewed from today – classical media. The main topics along the **institutional aspects** deal with course environments, base courseware technology, the development of teaching programmes, and the management of virtual education. Besides the core of web-based online and distance education, the field still deals with the more broad topic of

computer-based training. **Media themes** – computers, the web/internet, multimedia, and hypermedia – are the dominating not only peripheral clusters.

Eight years later in 2008, the interests have shifted. **Media** has become less dominant, although peripheral clusters deal with e.g. podcasting. Whereas the application and the development of tool and **(social) media usage competence** gained ground: multimedia usage skills, community technology (approaches and evaluation), and the improvement of **virtual content** play an important role. **Policy aspects** such as the development of knowledge training **strategies**, study/collaboration/instruction **practice**, and pedagogy **integration** now form important clusters. **Valorisation** (value generation, success, future ict) has gained ground. The **institutional** point of view is still reflected – distance education programmes and systems –, but especially the cluster formerly containing teaching and training frameworks now has opened up to put the **learner** (student) with peers and **professional learners** centre stage, involving process and service aspects. **Social effects** on education form a cluster, similarly **culture and interaction** have dissipated into the backbone structure.

To draw an overall conclusion, the field clearly can be asserted to be opening up to new areas, diversifying research of old clusters, and extinguishing certain terminology as interests shift and research has produced solutions for some of the underlying problems.

5 Outlook

Several extensions of this analysis would be useful to gain further insights: first of all, an authorship network analysis would complement the picture about the state of competence in the network to cover also rich professional aspects. Other than classical citation analysis, co-author graphs tend to disclose information about the nature of knowledge production rather than perception and uptake.

Furthermore, bipartite networks integrating both topical closeness and social relationships impose another unresolved challenge on the interested analyst.

Acknowledgement

This work has been financially supported by the European Union under the ICT programme of the 7th Framework Programme in the project ‘Stellar’ (grant agreement number 231 913).

References

Butts, C. (2007): sna. R package version 1.5

Butts, C.; Hunter, D.; Handcock, M. (2008): network. R package version 1.4-1

Erpenbeck, J.; Rosenstiel, L. (2003): Handbuch Kompetenzmessung, Schäffer-Poeschel, Stuttgart/Germany

Freeman, L.C. (1979): Centrality in Social Networks I: Conceptual Clarification, Social Networks, 1, 215-239.

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

Maechler, M. (2008): cluster. R package version 1.11.11

Rychen, D.S.; Salganik, L.H. (2003): Key Competencies for a Successful Life and a Well-functioning Society, Hogrefe & Huber, Göttingen/Germany

Salton, G.; Wong, A.; Yang, C.A. (1975): A Vector Space Model for Automatic Indexing, In: Communications of the ACM, 18 (11), 613–620

Temple Lang, D. (2004): Rstem. R package version 0.2-0

Wild, F. (2008): lsa: Latent Semantic Analysis. R package version 0.61

Wild, F.; Stahl, C. (2007): Investigating Unstructured Texts with Latent Semantic Analysis, In: Lenz & Decker (Eds.): Advances in Data Analysis, 383-390, Springer, Berlin/Germany

Wild, F.; Stahl, C.; Stermsek, G.; Neumann, G. (2005): Parameters Driving Effectiveness of Automated Essay Scoring with LSA, in: Proceedings of the 9th International Computer Assisted Assessment Conference (CAA), 485-494, Loughborough/UK