Computational analysis and mapping of ijCSCL content

Jacques Lonchamp

Abstract The purpose of this empirical study is to analyze and map the content of the *International Journal of Computer-Supported Collaborative Learning* since its inception in 2006. Co-word analysis is the general approach that is used. In this approach, patterns of co-occurrence of pairs of items (words or phrases) identify relationships among ideas. Distances based on co-occurrence frequencies measure the strength of these relationships. Hierarchical clustering and multidimensional scaling are the two complementary exploratory methods relying on these distances that are used to analyze and map the data. Some interesting findings of the work include a map of the key topics covered in the journal and a set of complementary techniques for investigating more specific questions.

Keywords CSCL · Content analysis · Co-word analysis · Hierarchical clustering · Multidimensional scaling · Topics · Themes

Introduction

The *International Journal of Computer-Supported Collaborative Learning* (*ijCSCL*) is defined on its Web site (ijcscl.org) as a "multidisciplinary journal dedicated to research on and practice in all aspects of learning and education with the aid of computers and computer networks in synchronous and asynchronous distributed and non-distributed groups." The journal appears quarterly since 2006. *ijCSCL* has published 121 articles containing about 1.5 million words as of the end of 2011. Most of the 24 issues are not organized around particular themes, except five of them that featured several articles on "flash themes" devoted to community-based learning, scripting, argumentation, evaluation methods, and tabletop computing. The empirical study reported in this article aims at providing a factual view of the CSCL research field through a computational analysis and mapping of this highly specialized text corpus.

J. Lonchamp (✉)
LORIA-Université de Lorraine, Campus Scientifique, BP239, 54506 Vandœuvre-lès-Nancy Cedex, France
e-mail: jloncham@loria.fr

There exist three main approaches for analyzing and mapping a research domain through a corpus of texts, either author-based or content-based. (1) In the first approach, co-citation analysis (White and Griffith 1981), two authors are associated if they are cited together, regardless of which of their work is cited. Co-citation analysis can be used to infer the intellectual structure of a field, its history and current front (e.g., White and McCain 1998). (2) In the second approach, co-authorship analysis (Price and Beaver 1966), two researchers are associated if they have written a paper in common. Co-authorship analysis is used to show the social network of a field (e.g., Liu et al. 2005). (3) The third approach, co-word analysis (Callon et al. 1983), is content-based. Co-word analysis may provide insights into the popularity of specific topics, the way topics relate to one another and the evolution of popularity of topics over time. It does not require any manual coding of the corpus, unlike other content-based approaches. The words that are used are extracted from the title, the keyword list, the abstract, or the full text of each article. Co-word analysis uses patterns of co-occurrence of pairs of items (words or phrases) to identify relationships among ideas. Distances based on co-occurrence frequencies are used to measure the strength of these relationships. Various clustering techniques relying on these distances can be used for analyzing and mapping the data. By comparing cluster maps for different time periods, the dynamics of a field can be detected. The present study adopts the full-text co-word analysis methodology. Hierarchical clustering and multidimensional scaling methods are used to analyze and map the data. The study was carried out with the aid of the WordStat software (www.provalisresearch.com).

The remainder of this paper is organized as follows. The next section discusses two key concepts behind co-word analysis: the "bag of words" representation of documents and the "word-word co-occurrence matrix". Then, the two complementary exploratory methods that were used in the study, hierarchical clustering and multidimensional scaling, are briefly outlined. The core of the article, in the following two sections, describes the methodology of the study and its main results. Some interesting findings include a map of the key topics covered in the journal and a set of complementary techniques for investigating more specific questions.

The approach

### The "bag of words" representation

The "bag of words" representation of documents (Zellig 1954) is a simplifying assumption used in domains like natural-language processing and information retrieval. In this model, a document is represented as a collection of "words," disregarding grammar and ordering. The term "word" may be given different meanings: token, token type, or higher-level construct. A token is simply the occurrence of a string in a text (including numbers, abbreviations, acronyms, etc.). For example, this sentence has seven tokens. A token type is a string that occurs one or more times in a text. Unlike the sentence above, this sentence has 13 tokens but only 12 types. There is also the possibility to consider higher-level constructs ("phrases") like n-grams (continuous sequences of n tokens), noun sequences (requiring the use of a grammatical tagger), or manually entered qualitative codes that categorize document content. Moreover, words are often pre-processed by natural-language processing techniques such as lemmatization (Beale 1987), stemming (Porter 1980), automatic spell correction, and exclusion. In many languages, words appear in different inflected forms. For example, in English, the verb "to learn" may appear as "learn," "learned," "learns," or "learning." The

base form, "learn", is called the lemma. Lemmatization is the task of finding the lemma of a given word form. The process is similar, while not identical, to the task of stemming, which removes affixes from a word and returns the stem (the largest common part shared by morphologically related forms). For instance, the word "better" has "good" as its lemma. This link is missed by stemming, as it requires a dictionary look-up. Exclusion is the process of suppressing token types that are extremely common and have little discriminative value, often called "stop words." There is not one definitive list of stop words. A list generally includes short function words (like "the," "is," "at," "which") and the most common words such as "want". Zipf's law (Zipf 1932) specifies that given some corpus, the frequency of any word is inversely proportional to its rank in the frequency table. In English, the most frequent type is about 10 % of the tokens and the most frequent 100 types are 52 % of the tokens. Most token types are very rare and can be safely ignored during content analysis. In English 50 % of the types occur once and 91 % of the types occur fewer than ten times. Figure 1 shows where in the rank-frequency distribution the most useful words for content analysis should be located.

Corpus matrices

A corpus of documents, where each document is considered as a "bag of words," is generally represented by two matrices. In the "word-document matrix" each entry gives the occurrence frequency of a particular word in a particular document, or some more sophisticated weight. In the "word-word co-occurrence matrix", each cell gives the frequency that two particular words co-occur. A co-occurrence happens every time two words appear in a same zone, which can be the whole document, a paragraph, a sentence, or a window of n consecutive words. Small zones, such as sentences, allow identifying idioms. Larger zones, such as paragraphs, are more appropriate to identify the co-occurrence of themes. Another usual weight is "term frequency weighted by inverse document frequency" (the product TF.IDF). It is based on the assumption that a word (term) is useful for determining the topic of a document if it appears in relatively few documents, but when it appears in a document it tends to appear many times. The TF part of the product can be normalized to prevent a bias towards longer documents, which may have a higher term count regardless of the actual importance of that term in the document. The IDF part of the product is generally obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient (Salton and McGill 1986).

Fig. 1 The most useful words for content analysis
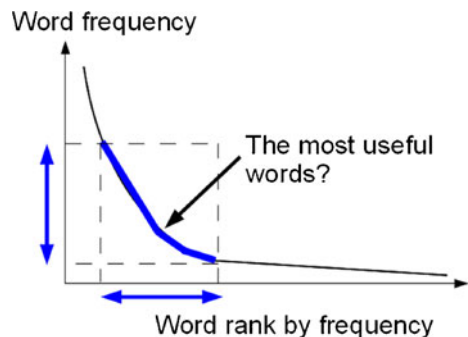
Hierarchical clustering

Two main approaches, clustering and dimensionality reduction, are used for analyzing and mapping the data contained in corpus matrices. In general terms, a cluster problem can be formulated by a set of objects and a distance function. The goal is to divide the object set into a number of sub-sets (clusters) that best reveal the structure of the object set. The most famous clustering methods are K-means (MacQueen 1967) and Hierarchical Clustering (HC) (Jardine and Sibson 1971). There is no clear consensus on which of the two produces better results. The present study uses agglomerative HC that does not require a pre-determined number of clusters (at the difference of K-means) and can give different solutions depending on the level-of-resolution used. The two criticisms against HC are that it is based on a local not undoable optimization and it is expensive in terms of computational and storage requirements. Agglomerative HC can be outlined in the following terms. Given a set of N items to be clustered, and a NxN distance matrix:

(1) Start by assigning each item to its own cluster (N clusters with a single item). (2) Find the closest pair of clusters and merge them into a single cluster, so that now there is one less cluster. (3) Compute distances between the new cluster and each of the old clusters. (4) Repeat steps 2–3 until all items are clustered into a single cluster of size N. The process is based on the notion of distance (or dissimilarity) between items in the initial matrix and distance between clusters during the algorithm. There exist many ways of measuring the similarity (or dissimilarity computed as 1-similarity) between items, that is the strength of the co-occurrence linkage in the present case. The more usual are Jaccard index ($C_{ij}/(C_i + C_j + C_{ij})$, where C_{ij} is the number of cases where both words occur, C_i and C_j the number of cases where one word is found but not the other) and cosine index (cosine of the angle between two vectors of values). The first one takes into account the presence/absence of a word but not joint absences (and therefore not fully word frequencies) at the difference of the second one. Similarly, there exist many ways of measuring cluster distance: the shortest distance from any member of one cluster to any member of the other cluster in "single-link clustering," the longest distance from any member of one cluster to any member of the other cluster in "complete-link clustering," and the average distance (weighted or un-weighted) from any member of one cluster to any member of the other cluster in "average-link clustering" (see Fig. 2).

A more advanced clustering technique is "second order clustering." In that approach, two words are close to each other, not necessarily because they co-occur but because they both occur in similar environments. For example, while "tumor" and "tumour" will probably never occur together in the same document, second order clustering may find them to be close because they both co-occur with words like "brain" or "cancer." Second order clustering also groups words that are related semantically such as "milk," "juice," and
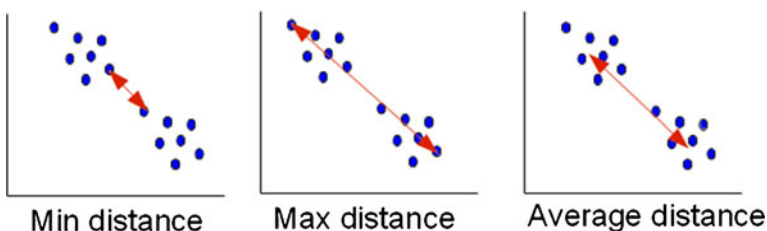

Fig. 2 Cluster distances

"wine" because of their propensity to be associated with similar verbs like "drink" or nouns like "glass" (Grefenstette 1994).

HC results are displayed by a tree of clusters, called "dendogram," which shows at what level of similarity any two clusters were joined. Determining a "good decomposition" into clusters is far from being obvious. The following rules of thumb may be helpful: (1) Highly correlated clusters are, by definition, near the bottom of the dendrogram. (2) As the agglomeration process occurs, increasingly dissimilar clusters are agglomerated; therefore, clusters should not include very long stems. (3) Clusters should preferably be defined at the same level of similarity ("fixed height branch cut"). So, when drawing a line at some chosen level, all stems that intersect that line indicate a cluster. In the example of Fig. 3, cutting the dendrogram at level 8 of dissimilarity gives five clusters ({4, 15, 1}, {2, 11, 6, 9, 14, 17, 0, 3, 10, 16}, {7, 12}, {8, 18, 19, 5}) and one isolated object (13). The size of the second cluster may appear too large. Another possible solution is to cut lower, at level 4 of dissimilarity, which gives six highly correlated clusters ({4, 15}, {2, 11, 6}, {9, 14, 17}, {0, 3}, {10, 16}, {8, 18, 19}) and five isolated objects (1, 7, 12, 5, 13). The positive side is that the large cluster is now divided into four subgroups with higher internal cohesion. The negative side is that some peripheral clusters are lost, such as {7, 12}, and that more isolated elements appear. The "fixed height branch cut" strategy is not always ideal (Langfelder et al. 2008). A multi-step approach can help to identify nested clusters. In the first step, the cut is placed at a level where a set of significant "peripheral" clusters appears around one or a few "big agglomerates". In the following steps only these big agglomerates are further decomposed. In Fig. 3, this could lead to a solution including {4, 15, 1}, {7, 12}, and {8, 18, 19, 5}, resulting from the first decomposition, and {2, 11, 6}, {9, 14, 17}, {0, 3}, {10, 16}, resulting from the second decomposition. However, in all cases, a "good decomposition" basically is a decomposition that can be clearly interpreted.

In general, HC is no longer considered as the "best" clustering approach for documents. When repeatedly executing a given algorithm with many document sets in which each document can be pre-classified into a single class, the F-measure shows that bisecting K-means and regular K-means for instance perform better than HC in terms of accuracy of the clustering results (Steinbach et al. 2000). Because of the probabilistic nature of how words are distributed, any two documents may share many of the same words. Thus, in 5–30 % of cases (for the different document sets of the above mentioned study), nearest neighbours that

Fig. 3 A dendogram
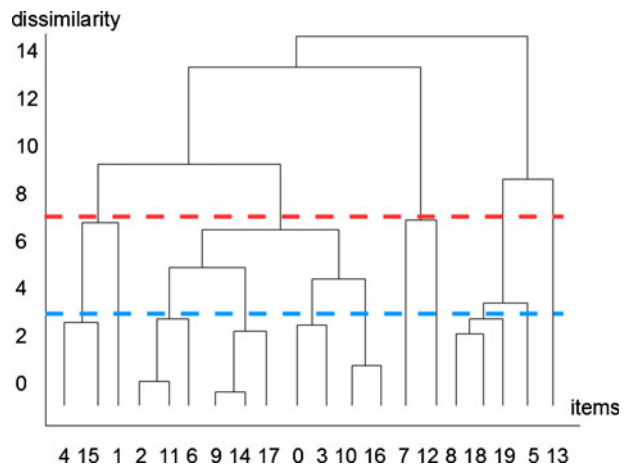
belong to different classes are put in the same cluster, even at the earliest stages of the clustering process. Because of the way HC works, these "mistakes" cannot be fixed once they happen. It is important to note that HC is used very differently in this study, that is, in an exploratory mode with a single corpus. The many parameters (pre-processing, item distance, cluster distance, similarity level) are fine-tuned in an iterative manner by the specialist of the domain until finding the more "meaningful solution". Local optima can be avoided by having multiple trials. In that specific perspective, HC remains valuable due to its simplicity and flexibility.

Multidimensional scaling

The second fundamental method in content analysis is multidimensional scaling (MDS) (Kruskal and Wish 1978). It is part of the dimensionality reduction approach, together with eigenvalue/eigenvector decomposition (Francis 1961), factor analysis (Spearman 1950), latent semantic analysis (Deerwester et al. 1990), and self-organizing maps (Kohonen 1982). All these methods aim at deriving useful representations of high dimensional data. MDS attempts to find the structure in a set of dissimilarity measures among objects. This is accomplished by solving a minimization problem such that the distances between points in the target low-dimensional space match the given dissimilarities as closely as possible. If the dimension is chosen to be 2 or 3 one may plot the points to obtain a visualization of the similarities among the objects (2D or 3D map). MDS is a numerical technique that iteratively seeks a solution and stops computation when an acceptable solution has been found or after some pre-specified number of attempts. It actually moves objects around in the space defined by the requested number of dimensions, and checks how well the distances among objects can be reproduced by the new configuration. In other terms, it evaluates different configurations with the goal of maximizing a "goodness-of-fit" measure or "stress" measure. The stress value of a configuration is measured based on the sum of squared differences between the reproduced distances and the given distances. The smaller the stress value, the better is the fit of the reprod/awuced distance matrix to the given distance matrix. The R-square value determines what proportion of variance of the scaled data can be accounted for by the MDS procedure. An R-square of 0.6 is sometimes proposed as the minimum acceptable level. The strength of MDS is that it can be used to analyze any kind of distance matrix. MDS suffers from several drawbacks: it is slow for large data sets, it can get stuck on a local minimum, and there are no simple rules to interpret the nature of the resulting dimensions. HC and MDS are complementary methods that can be used with the same distance matrix (Kim et al. 2000). HC is good for understanding the divisive process of clustering while MDS is good for comparing different solutions by displaying the relative positions of words and clusters, and their distances.

The data

The first stage of the content analysis process, data collection, is the construction of the corpus. The present study uses three corpora. The first one, the "full text corpus," encompasses the 121 articles appeared in the 24 first issues of *ijCSCL*, as retrieved from the publisher web site (www.springer.com). The second one, the "abstracts corpus," contains the abstracts of all the articles. The Web site of *ijCSCL* requires an informative abstract of 100–250 words without undefined abbreviations or unspecified references. Each abstract has been stored in a separate text file. The third one, the "keywords corpus," contains the

keyword lists of all the articles. Authors freely choose three to ten keywords or short phrases. Each keyword list has been placed in a separate text file.

The second stage of the content analysis process involves data extraction and standardization. From the three corpus respectively 1,420,870 22,906 1335 tokens and 29,178 3140 490 token types were extracted. This corresponds to an average length of 11,743 tokens by article, 189 tokens by abstract and 11 tokens by keyword list. Table 1 gives the 20 most frequent token types in the three corpora. The first two lists contain many words with little semantic value and inflected forms such as "learning" or "students." Without surprise, the "keywords corpus" gives fewer words with little discriminative value.

The third stage is lemmatization and automated exclusion of stop words. Figure 4 gives examples of substitutions and an excerpt of the list of stop words. The lemmatization algorithm and the standard list of stop words provided by the software were used. Table 2 gives the resulting 20 most frequent token types from the three corpora. Effects of lemmatization and exclusion are easy to see: "learn" instead of "learning," "student" instead of "students" and less non-discriminative words.

However, some additional cleaning is still required. For example, "base" token type comes from expressions like "computer-based" and has little semantic interest. A "word-in-context window" helps in analyzing these cases by displaying all occurrences of a word together with the textual environment (phrase or paragraph) in which they occur. Moreover, most of these high ranked words appear in nearly all documents of the full text corpus (see "NO.DOCS" column in Table 2). Logically, the fact is less apparent in the two other corpora

Table 1 Most frequent words

Full text corpus		Abstracts corpus		Keywords corpus	
Token type	Freq.	Token type	Freq.	Token type	Freq.
The	83,839	The	1,349	Learning	84
Of	51,359	Of	1,017	Collaborative	42
And	40,661	And	783	Analysis	38
To	34,910	In	635	Knowledge	29
In	34,438	To	605	Collaboration	25
A	31,556	A	553	CSCL	25
That	15,400	That	302	Computer	25
For	14,091	For	269	Of	21
Is	13,602	Learning	258	Interaction	18
As	10,994	This	231	Education	17
Learning	10,876	Students	213	Based	15
On	10,645	Is	201	Argumentation	14
With	8,671	On	191	Social	13
This	8,596	We	188	Supported	13
S	8,500	As	173	Design	13
Students	8,191	Knowledge	158	Group	12
Be	7,295	With	151	Online	12
Are	7,257	Collaborative	135	Discussion	11
By	7,210	An	132	Problem	11
We	6,429	Are	122	Practice	11

Fig. 4 Examples of substitutions and excerpt of the list of stop words

Original	Substituted
EVALUATIONS	EVALUATION
GENERALIZED	GENERALIZE
GROUPS	GROUP
IMPROVING	IMPROVE
INTEGRATED	INTEGRATE
INTERPRETATIONS	INTERPRETATION
ISSUES	ISSUE
LARGER	LARGE
LEARNING	LEARN
LEARNERS	LEARNER
LEVELS	LEVEL
LOGS	LOG
MEANINGFUL	MEANING
MODELS	MODEL
MONITORING	MONITOR
OBSERVED	OBSERVE
OPENS	OPEN
ORGANIZED	ORGANIZE
ORIENTED	ORIENT
PARAMETERS	PARAMETER
PEDAGOGICAL	PEDAGOGIC
PERFORMING	PERFORM
PROPOSED	PROPOSE
PROVIDED	PROVIDE

```
O
OBVIOUSLY
OF
OFF
OFTEN
OH
OK
OKAY
OLD
ON
ONCE
ONE
ONES
ONLY
ONTO
OR
OTHER
OTHERS
OTHERWISE
OUGHT
OUR
OURS
OURSELVES
OUT
OUTSIDE
OVER
OVERALL
OWN
```

whose elements contain fewer words. It is where TF.IDF weighting is useful. Table 3 gives the 20 selected token types with the higher TF.IDF weights in the three corpora. There is a radical change in the full text corpus list in which all the items are changed. Specialized terms such as "script", "wiki", "argumentation", "ontology", or "tabletop" appear. The change is less important in the abstracts corpus list in which there is a mix of general and specialized terms. There is nearly no change in the keywords corpus list. This shows that TF.IDF weighting works well when there is a large collection of words in which specialized terms that are highly representative of specific themes can be found. When the collection size diminishes, TF.IDF weights and occurrence frequencies tend to be more similar. In the rest of the study, the full text corpus is mostly used because it gives better results for thematic analysis.

The concept of "phrase" (continuous sequence of tokens) is also interesting. Table 4 gives the ten highly ranked phrases that appear in at least two documents in the full text and abstracts corpora (weighted by TF.IDF). Frequencies of phrases are obviously lower than frequencies of isolated words. Nevertheless, phrases may represent important elements for topic characterization. So, the final list of candidate items in the study includes both words and phrases. The weight of phrases, which are generally "bigrams," has been doubled.

In the last stage of the content analysis process, the clustering method and the size of the word-word co-occurrence matrix, that is, the number of words and phrases that will be analyzed, has to be chosen. The clustering method that is used in the study is average-linkage HC with co-occurrence established at the paragraph level. The best distance

Table 2 Most frequent selected words

Full text corpus			Abstracts corpus			Keywords corpus	
Token type	Freq.	NO.DOCS	Token type	FREQ.	NO.DOCS	Token type	Freq.
Student	10,360	119	Student	240	72	Learn	85
Learn	8,122	120	Learn	205	94	Collaborative	42
Group	7,332	120	Knowledge	154	54	Analysis	38
Knowledge	5,612	120	Group	151	59	Knowledge	29
Computer	4,823	120	Study	144	78	Collaboration	25
Activity	3,937	119	Design	101	55	CSCL	25
Design	3,800	119	Support	96	62	Computer	25
Process	3,480	119	Activity	92	45	Support	24
Discussion	3,409	121	CSCL	91	43	Education	20
Interaction	3,404	119	Analysis	90	50	Interaction	18
Education	3,252	120	Process	89	42	Group	15
Analysis	2,993	120	Interaction	86	46	Base	15
Problem	2,985	118	Collaborative	84	69	Practice	15
Work	2,836	121	Collaboration	77	37	Discussion	14
Tool	2,588	115	Discussion	74	30	Argumentation	14
Base	2,560	121	Tool	70	32	Community	14
Model	2,499	115	Education	67	39	Social	13
Teacher	2,467	109	Technology	65	31	Tool	13
Collaborative	2,414	117	Base	62	43	Script	13
System	2,334	114	Practice	59	31	Problem	13

definition is chosen through R square maximization, among four possibilities: Jaccard and cosine, first and second order. As the same matrix is used for HC and MDS, its size must be not too large, because huge maps are hard to draw and analyze. Other content analysis studies also use matrices of limited sizes. For example, Ding et al. (2001) have selected 240 items from 3 227 unique keywords coming from 2 012 articles for mining information retrieval research. Three word-word co-occurrence matrices have been tested: a 200×200 matrix and a 120×120 matrix extracted from the full text corpus and a 120×120 matrix extracted from the abstracts corpus.

The results

Elementary statistics

Occurrence frequencies in the full text corpus can be used to answer some simple questions, such as the most popular tools or persons. Answers are given in Table 5. It is interesting to note, for instance, the surprisingly low importance of social networking tools, such as Facebook, when compared with chat and wiki tools that are widely studied. Specific tools are rarely mentioned, with the exception of Knowledge Forum, which remains the "flagship" CSCL tool. Concerning persons, the measure is a basic count of names in text as well as references, which excludes articles where authors cite their own works. The results are

Table 3 Selected words with the higher TF.IDF weights

Full text corpus			Abstracts corpus			Keywords corpus		
Token type	Freq.	TF.IDF	Token type	Freq.	TF.IDF	Token type	Freq.	TF.IDF
Wiki	795	505.3	Student	240	57.0	Learn	85	27.8
Tabletop	396	467.2	Knowledge	154	55.2	Analysis	38	24.8
Facebook	248	398.2	Group	151	49.4	Knowledge	29	21.5
Annotation	507	396.3	Discussion	74	45.9	Collaborative	42	20.7
Script	1,212	360.5	Argumentation	48	45.0	Collaboration	25	18.5
Game	669	360.4	Script	40	43.3	CSCL	25	18.5
Argumentation	1,329	358.7	CSCL	91	42.8	Support	24	18.3
Chat	914	327.7	Activity	92	42.3	Computer	25	18.0
Teleport	197	316.3	System	54	42.2	Community	14	17.3
Ontology	381	289.8	Process	89	41.8	Education	20	16.1
Tutor	752	288.6	Collaboration	77	40.5	Interaction	18	15.8
Instructor	604	277.6	Tool	70	40.4	Practice	15	15.6
Graph	432	268.0	Online	56	40.4	Model	12	14.9
Message	1,144	264.9	Technology	65	38.4	Discussion	14	14.1
Phase	1,122	259.8	Interaction	86	37.8	Argumentation	14	14.1
Diagram	492	247.5	Model	55	37.7	Tool	13	14.1
Argument	1,550	245.7	Collaborative	84	36.9	Problem	13	14.1
MetaCognitive	451	237.4	Learn	205	36.8	Base	15	14.0
Debate	574	225.3	Classroom	51	35.8	Group	15	14.0
Digalo	181	224.0	Design	101	35.4	Social	13	13.5

consistent with the list of key members given by Kienle and Wessner (2006) in their study of CSCL community membership. However, it can be noticed in Table 5 a few names of historical and more recent pioneers, such as Lev Vygotsky, David and Roger Johnson, Marlene Scardamalia and Carl Bereiter, who do not appear in Kienle and Wessner's list.

Table 4 Phrases with higher TF.IDF weights

Full text corpus				Abstracts corpus			
Phrases	Freq.	NO.DOCS	TF.IDF	Phrases	Freq.	NO.DOCS	TF.IDF
Knowledge forum	300	18	248.3	Collaborative learning	50	35	26.9
Metacognitive activities	127	4	188.1	Knowledge construction	16	8	18.9
Dialogue acts	106	3	170.2	Knowledge building	13	8	15.3
Floor control	142	8	167.5	Online discussions	11	5	15.2
Knowledge domain	91	2	162.1	Collaboration scripts	10	4	14.8
Peer tutor	90	2	160.4	Collaborative knowledge	13	9	14.7
Automatic coding	90	3	144.5	Learning processes	11	7	13.6
Group members	353	49	138.6	Problem solving	10	7	12.4
Collaboration scripts	204	27	132.9	Case study	11	11	11.5
Group formation	117	9	132.0	Knowledge forum	7	3	11.2

Table 5 Simple counts in the full text corpus

Tools	Freq.	Persons	Freq. (without self citations)
Chat	914	Suthers	411
Wiki	795	Dillenbourg	359
Web	717	Stahl	292
Forum	537	Fischer	268
Tabletop	396	Scardamalia	249
Video	310	Bereiter	243
Internet	309	Weinberger	224
Knowledge forum	300	Baker	210
Mail	271	Koschmann	194
Facebook	248	Johnson	190
Whiteboard	206	Vygotsky	172
Database	190	Kirschner	146

This could be interpreted as an indicator that *ijCSCL* articles are well-grounded in a wide body of research and strong theoretical foundations.

By comparing occurrence frequencies of related words some other interesting characteristics can be observed. As a first example, synchronous and asynchronous modes of interaction appear roughly equally mentioned in articles: "synchronous" (402 occurrences) and "synchronously" (12) are slightly more frequent than "asynchronous" (346) and "asynchronously" (15). As a second example, frequencies show that much more occurrences of words designate learners than teachers: "student" (10 360), "learner" (1 560), "peer" (997), "child" (948), and "pupil" (235) are far more frequent than "teacher" (2 467), "tutor" (752), "instructor" (604), and "educator" (64). The full text corpus, and probably the CSCL field, focuses more on the learner side than on the teacher side.

Global thematic analysis

Most dendograms produced by HC are difficult to analyze, because clusters are defined at all levels of dissimilarity. The solution of selecting a large number (like a hundred) of highly consistent clusters by cutting the tree at a low level of dissimilarity is not satisfying, because it means more or less associating one cluster to each article. To have a chance of characterizing general topics the number of clusters must be much lower. An interesting decomposition has been obtained empirically with the following characteristics: full text corpus, 120×120 matrix, 25 clusters, Jaccard second order distance, R2=0.56. Figure 5 shows the corresponding 2D map: clusters have different colors (apparent in the online version of the article) and circles are proportional to the weights. This map contains a "big cluster" (circled in Fig. 5) in the middle that is difficult to interpret at this level. At the opposite, many small "peripheral" clusters are easy to interpret.

The most obvious are given in the following list with their higher frequency word in bold face (they are boxed in Fig. 5):

- CL1 = {**script**, collaboration script, macro (script), Dillenbourg, platform}
- CL2 = {**uptake**, contingency (graph), event, node}
- CL3 = {**map**, ontology, Digalo, moderator, discussant}
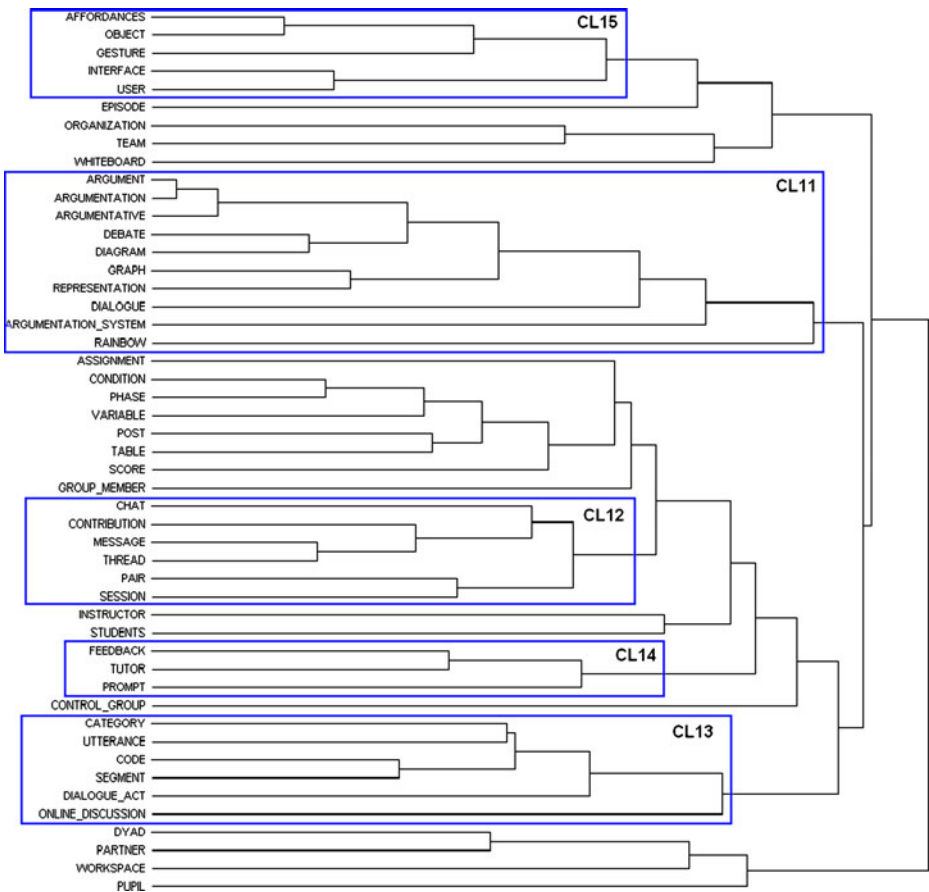- CL4 = {**tabletop**, touch, interference}

Fig. 5 2D map of the full text corpus (first level analysis)

- CL5 = {**mathematic**, mathematical proof},
- CL6 = {**metacognitive**, metacognitive activity}
- CL7 = {**knowledge building**, knowledge creation, scaffold, community, Knowledge Forum, Scardamalia, Bereiter, collective, inquiry, discourse, portfolio, database}
- CL8 = {**argumentation sequence**, counterargument}
- CL9 = {**wiki**, page, blog, Coweb}
- CL10 = {**floor control**, group formation}

It is interesting to note that the five "flash themes" defined by *ijCSCL* editors can be easily related to five of these clusters: community-based learning to CL7, scripting to CL1, argumentation to CL8, evaluation methods (more precisely media independent interaction analysis) to CL2, and tabletop computing to CL4. The five other clusters may be interpreted in the following way: CL3 can be linked to the theme of collaborative mapping, CL5 to the dominant application domain (mathematical applications), CL6 to metacognitive support, CL9 to collaborative writing (wikis), and CL10 to collaborative services.

In the dendogram of the big central cluster (see Fig. 6) five easily interpretable sub-clusters that stay at similar levels of dissimilarity can be detected (they are boxed):

- CL11 = {**argument**, argumentation, argumentative, debate, diagram, graph, representation, dialogue, argumentation system, rainbow (method)}
- CL12 = {**message,** thread, pair, session, contribution, chat}
- CL13 = {**code,** category, dialogue act, online discussion, utterance, segment}
- CL14 = {**feedback**, prompt, tutor}
- CL15 = {**object,** affordances, user, interface, gesture}

Figure 5 shows that CL8 and CL11 are located very close. As both deal with argumentation, they can be merged. CL12 and CL13 are about conversation analysis with a slightly different orientation in terms of message structure for the former (thread, pair, session) and in terms of

Fig. 6 Dendogram of the central cluster (second level analysis)

message content for the latter (code, category, dialogue act). Finally, CL14 is about monitoring and CL15 can probably be interpreted in terms of technical object affordances and user interface issues.

As is frequently the case with MDS, dimension interpretation is not straightforward. A visual inspection of Fig. 5 may suggest that:

(1) Clusters in the upper part of the map (see Fig. 7a) deal with applications (mathematics, argumentation, tabletop computing, shared writing, collaborative mapping, community based learning…).

(2) Elements in the lower part (see Fig. 7b) predominantly deal with technological and organizational issues (scripting, floor control, feedback, metacognitive support, group formation, efficacy, adaptive…).

(3) Items related to interaction analysis issues mostly stay in the central part of the map as shown by Fig. 7c (uptake, contingency, message, thread, pair, session, segment, code, category, dialogue act, utterance…).

Table 6 summarizes the main findings of this global thematic analysis.

Fig. 7 a Applications. b Technological and organizational issues. c Interaction analysis

Table 6 Summary of the global thematic analysis

Categories	Themes	Keywords
Applications	Argumentation	**Argument**, argumentation, argumentative, debate, diagram, graph, argumentation system, rainbow (method), argumentation sequence, counterargument…
	Community-based learning	**Knowledge building**, knowledge creation, scaffold, community, Knowledge Forum, Scardamalia, Bereiter, collective, inquiry…
	Collaborative writing	**Wiki**, page, blog, Coweb…
	Collaborative mapping	**Map**, ontology, Digalo, moderator…
	Tabletop computing	**Tabletop**, touch, interference…
	Mathematical applications	**Mathematic**, mathematical proof…
Technical and organizational issues	Scripting	**Script**, collaboration script, macro (script), Dillenbourg…
	Monitoring	**Feedback**, prompt, tutor…
	Affordances and user interface issues	**Object,** affordances, user, interface, gesture…
	Metacognitive support	**Metacognitive**, metacognitive activity…
	Collaborative services	**Floor control**, group formation…
Interaction analysis	Cross media analysis	**Uptake**, contingency (graph), event, node…
	Conversation structuring	**Message,** thread, pair, session, contribution…
	Message content	**Code,** category, dialogue act, utterance…

Complementary thematic analysis

The co-word approach may also be used for exploring more specific questions and particular semantics fields. Two examples are given. In the first one, HC and MDS allow characterizing the main issues related to the general concept of knowledge. First, all phrases including "knowledge" are retrieved in the full text corpus. Those with a TF.IDF weight greater than 25 are selected (see Table 7).

Then, clustering analysis is performed on this set of phrases. Figure 8 shows the resulting 2D map (Jaccard second order, nine clusters, R2=0.713) in which clusters with a single element are not drawn.

It is interesting to note the very high R2 square value. The resulting 2D map shows five clusters that are easy to interpret:

- KW1 = {**Knowledge Forum**, collaborative knowledge, collective knowledge, knowledge advancement} about group knowledge
- KW2 = {**prior knowledge**, knowledge acquisition} about individual knowledge (it could be extended with the term individual knowledge that is closely located)
- KW3 = {**tacit knowledge**, explicit knowledge} about the classical tacit/explicit duality
- KW4 = {**metacognitive knowledge**, factual knowledge} about the different levels of knowledge
- KW5 = {**knowledge construction**, knowledge management} about knowledge processes (could perhaps be extended with knowledge representation)

It is suggested to group the remaining keywords, all located at the bottom of the map, into a single cluster about categories of knowledge KW6 = {**conceptual knowledge**, procedural knowledge, design knowledge, intuitive knowledge}.

Table 7 Highly weighted phrases about knowledge

Phrases	Freq.	NO.DOCS	TF.IDF
Knowledge forum	299	18	247.4
Metacognitive knowledge	65	4	96.2
Tacit knowledge	79	9	89.2
Prior knowledge	142	29	88.1
Knowledge construction	257	57	84
Knowledge share	116	25	79.4
Collaborative knowledge	168	49	66
Knowledge advancement	60	10	65
Conceptual knowledge	78	18	64.5
Collective knowledge	71	15	64.4
Individual knowledge	75	27	48.9
Knowledge acquisition	67	23	48.3
Explicit knowledge	24	2	42.8
Design knowledge	31	7	38.4
Knowledge management	32	11	33.3
Intuitive knowledge	21	4	31.1
Knowledge representation	25	9	28.2
Factual knowledge	19	4	28.1
Procedure knowledge	23	9	26

In the second example, clustering is not used. The approach only considers distances among concepts, through "proximity plots." The question is to evaluate the empirical validity of a given assertion. Stahl (2002) has proposed four themes for "thinking about CSCL" that are "collaborative knowledge building, group and personal perspectives, mediation by artifacts and interaction analysis." The following three proximity plots, built with Jaccard first order distance on the full text corpus, are sufficient to evaluate the strength of that complex assertion and to suggest some additional themes. In the proximity plot of the "knowledge" word (see Fig. 9) associations with "collaborative," "building", "construction",
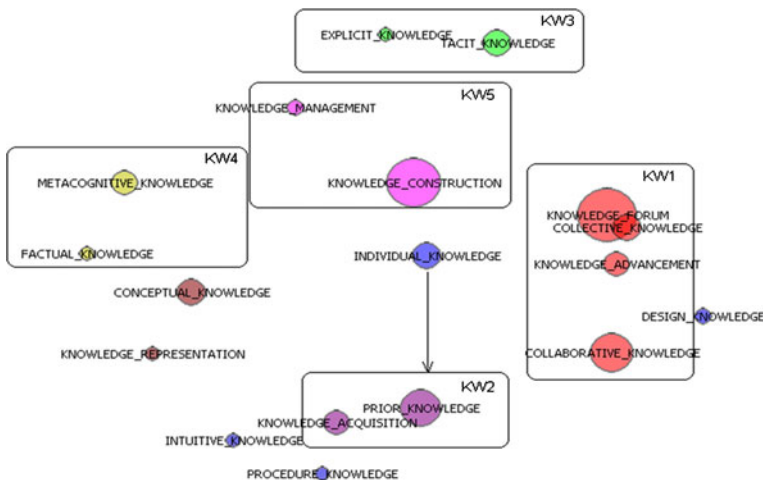
Fig. 8 "Knowledge" semantic field

"group", "individual" are all in the top ten proximities with Jaccard coefficients between 0.106 and 0.209 (showing that of all paragraphs containing either one of these words, between 10 and 20 % contain both words). This strongly supports the first part of the assertion about "collaborative knowledge building, group and personal perspectives".

In the proximity plot of the "artifact" word (see Fig. 10), the association with "mediate" is at the fourth rank showing the strength of the "mediation by artifacts" theme. The high rank of "property" and "affordances" could suggest an additional major theme about "artifact possibilities for collaborative learning." However, the associations are much weaker than in the previous case, with Jaccard coefficients between 0.04 and 0.05.

Finally, in the proximity plot of the "analysis" word (see Fig. 11) the association with "interaction" comes just after the trivial association with "data", with a Jaccard coefficient near 0.13. It might be worth considering also the larger theme of "process analysis" besides "interaction analysis", as suggested at the third rank.

These kinds of focused explorations are easy to perform as a complement to the global thematic analysis of the corpus and may have a great interest for specific research works.

Document clustering

The co-word approach applied to a large corpus of documents may help finding quickly groups of documents associated to the same themes. HC and MDS techniques are applied with a distance between documents computed from the document-word matrix. Figure 12 shows the whole 2D map obtained from a 121×200 document-word matrix by using cosine distance (60 clusters, R2=0.121). Figure 13 gives an excerpt of the corresponding dendogram focusing on the two clusters that are shown in rectangular boxes in Fig. 12.

The existence of strong relationships among the corresponding articles is easy to verify. In the first cluster, all articles deal with the issue of learning into communities:

- Fischer et al. (2007) consider community-based learning at the university level,
- Hung et al. (2008) examine the differences between para-communities (such as online communities) and schools,
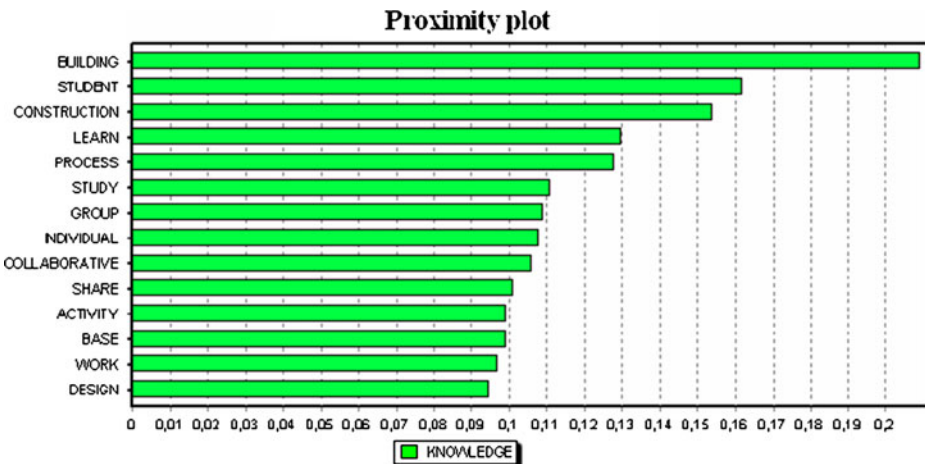- Nett (2008) analyzes a community of practice among tutors.

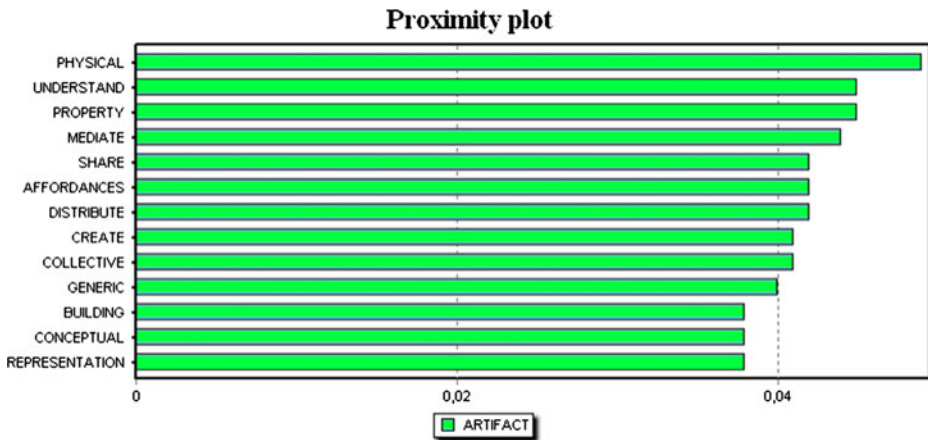Fig. 9 "Knowledge" proximity plot

Fig. 10 "Artifact" proximity plot

In the second cluster, all articles are about, or related to, collaborative learning with wikis:

- "A systemic and cognitive view on collaborative knowledge building with wikis" (Cress and Kimmerle 2008),
- "The right tool for the wrong task?" (Lund and Rasmussen 2008), which studies the relationship between a task and the tool (a wiki) that learners pick up, appropriate, and transform in order to make them serve their purpose,
- "Wiki-supported collaborative learning in primary education" (Pifarré and Staarman 2011),
- "Web 2.0: Inherent tensions and evident challenges for education" (Dohn 2009), which discusses discrepancies between the "practice logics" of Web 2.0 (including wiki usage) and educational practices,
- "Wikis to support the 'collaborative' part of collaborative learning" (Larusson and Altermann 2009),
- "The logic of wikis: The possibilities of the Web 2.0 classroom" (Glassman and Kang 2011).
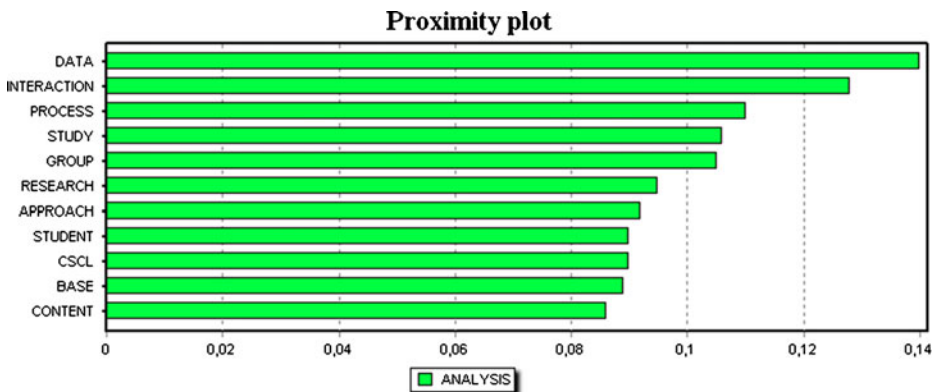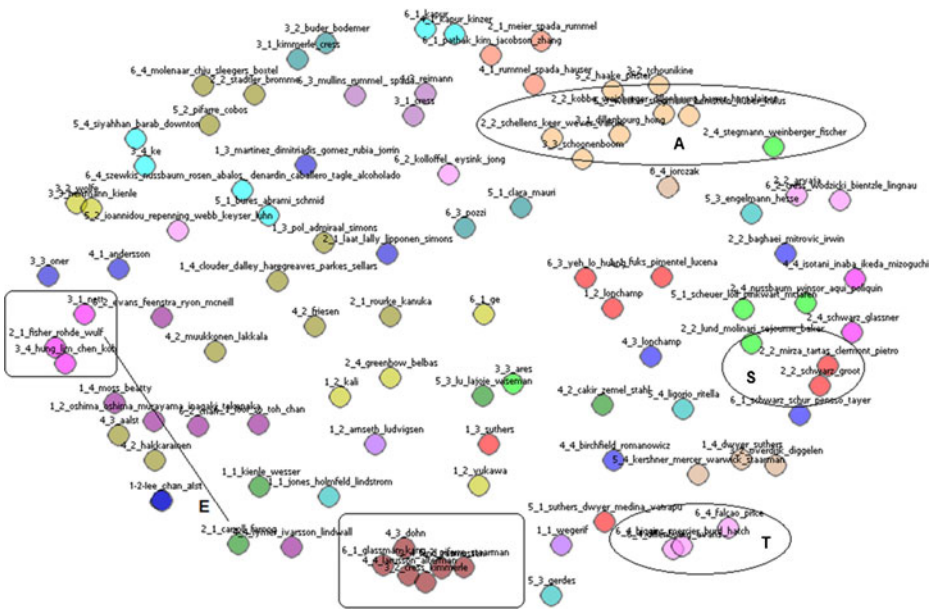

Fig. 11 "Analysis" proximity plot

Fig. 12 Document clustering (each cluster is assigned a specific color)

ijCSCL flash themes are also easy to locate in the map. In the best case a flash theme corresponds exactly to one cluster. It is the case of the "tabletop" flash theme shown in the oval box with the letter T in Fig. 12. In the intermediary cases, the papers of a theme are distributed into two close clusters. It is the case of the "scripting" flash theme (letter S in Fig. 12) and the "argumentation" flash theme (letter A in Fig. 12). The worst case is the "evaluation" flash theme whose two papers are not so close in the map (related by line E in Fig. 12). This probably reflects the fact that the definition of this last theme is less precise than the previous ones.

Thematic evolutions

Thematic evolutions of research fields can be studied by dividing a large corpus according to consecutive time periods and by contrasting content analysis results during these periods (e.g., Ding et al. 2001). The history of *ijCSCL* is rather short: it is impractical to divide it in

Fig. 13 Document dendogram excerpt
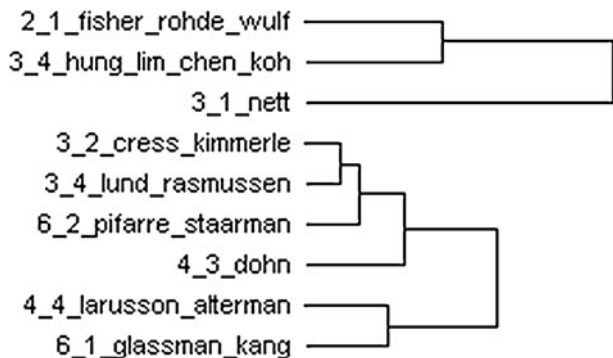
more than two periods of three years. Table 8 shows a comparison between word frequencies that were observed in full text during the second period (FREQ. column) and what was expected on the basis of the first period (Expected column). Ten high positive deviations are given on the left side of the table and ten high negative deviations are given on the right side, sorted by decreasing frequencies.

These results may suggest two topics of increasing popularity during the second period when compared with the first period:

- "CSCL in the classroom", evoked by the terms "teacher", "classroom", "peer", "child", and "inquiry".
- The various issues related to "CSCL support", with the terms "process", "scaffold", "representation", and "feedback".

Similarly, two themes of decreasing popularity may be suggested:

- "Basic communicative activities", evoked by the terms "discussion", "communication", "community", "note", and "text".
- "Argumentative activities", with "argument" and "argumentation".

These interpretations should be considered with great caution because a few articles may strongly impact word counts for a given period. For illustrating that phenomenon, Fig. 14 shows the chronological evolution of the count of the word "argumentation" into *ijCSCL* articles (if issues are ordered chronologically it is not the case of the papers in each issue, but the figure is nevertheless illustrative).

Conclusion

Co-word analysis is used in this article as a general approach for investigating the content of *ijCSCL* since its inception in 2006, with the purpose of giving a factual view of the CSCL research field as reported by the leading journal in the field. In that approach, hierarchical clustering and multidimensional scaling—two complementary exploratory methods—play a central role. They use graphical representations to display models. As highlighted by Edwards (1995) "in contrast to most other types of statistical graphics, the graphs do not display data, but rather an interpretation of the data, in the form of a model." This

Table 8 Frequency evolutions

| Words	Freq.	Expected	Dev.	Words	Freq.	Expected	Dev.
Process	2,195	1565.7	40.20 %	Design	1,693	2489.5	−32.00 %
Teacher	1,717	841.4	104.10 %	Discussion	1,607	2002.7	−19.80 %
Classroom	943	561.7	67.90 %	Model	1,202	1,492	−19.40 %
Post	759	326.4	132.60 %	Tool	1,142	1485.5	−23.10 %
Peer	750	300.3	149.70 %	Communication	719	919.5	−21.80 %
Scaffold	736	238.5	208.50 %	Argument	575	1064.8	−46.00 %
Representation	750	438	71.20 %	Argumentation	555	843.6	−34.20 %
Child	684	289.5	136.30 %	Community	464	1182.9	−60.80 %
Inquiry	639	386	65.50 %	Note	447	773.1	−42.20 %
Feedback	413	249.4	65.60 %	Text	447	722.1	−38.10 %
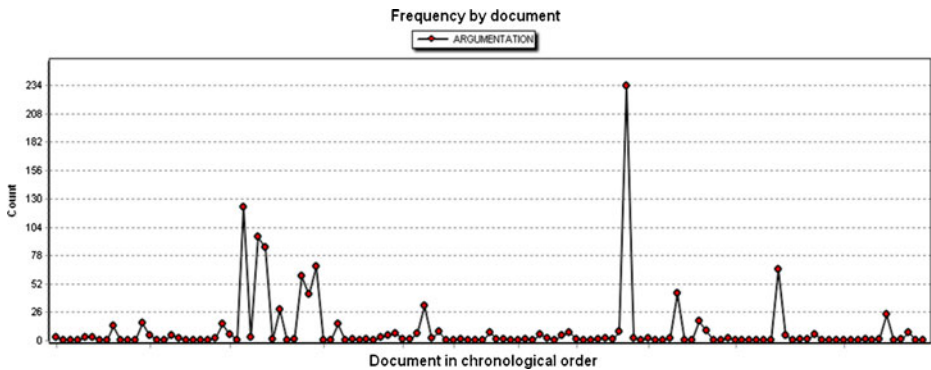
Fig. 14 Chronological evolution of the "argumentation" word count

interpretation by the machine depends on multiple parameters that must be empirically chosen by the analyst like the item frequency definition, the co-occurrence definition, the distance definition, the dissimilarity cutting level, and so forth. At the end of the process, the human interpretation of the graphical model adds another layer of subjectivity. So, the expression "factual view" used above when defining the purpose of this study is by no means to be understood as "objective view". It only means that it is elaborated from statistical facts.

The thematic analysis summarized in Table 6 is the main result of the work. "Interaction analysis", "CSCL technical and organisational issues" and "CSCL applications" are the three high level thematic categories that emerge. Each of them is further refined into three to six more focused themes characterized by a list of keywords. The analysis suggests five dominant themes in the "CSCL technical and organisational issues" thematic category that are entitled "scripting", "monitoring", "affordances and user interface issues", "metacognitive support" and "collaborative services". The "interaction analysis" thematic category reflects three levels of analysis, namely "message content", "conversation structuring" and "cross media analysis". Finally, the "CSCL applications" thematic category reveals six dominant themes that are "argumentation", "community-based learning", "collaborative writing", "collaborative mapping", "tabletop computing" and "mathematical applications". Another interesting finding is that co-word analysis is also effective for many focused exploratory tasks that can be of great interest for researchers, such as: (1) the exploration of document proximity for finding quickly groups of documents associated to a given theme in a large corpus, (2) the exploration of particular semantics fields, as exemplified by the study of the main research questions related to the general concept of knowledge, (3) the examination of specialized issues, as exemplified by the evaluation of an assertion about CSCL, and (4) the discovery of thematic evolutions and trends, with the condition of having a sufficiently large corpus of documents published over a long period of time, which is not yet fully the case for the *ijCSCL* corpus.

This work could be extended to a larger corpus by including articles in other journals explicitly mentioning the CSCL field and papers from the biannual International Conference on CSCL (from 1995). It would be interesting to see if this larger corpus would reveal additional topics and evolution trends.

In the future, scientific journals could provide as a service access to their full text corpus. Users would be supported for interactively applying content analysis techniques, like those

presented in this article, for gaining insights into specific patterns and trends, exploring particular fields, and retrieving documents by thematic proximities.

References

Beale, A. D. (1987). Towards a distributional lexicon. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 149–162). London: Longman.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Cress, U., & Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning, 3*(2), 105–122.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Ding, Y., Chowdhury, G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management, 37*(6), 817–842.

Dohn, N. (2009). Web 2.0: Inherent tensions and evident challenges for education. *International Journal of Computer-Supported Collaborative Learning, 4*(3), 343–363.

Edwards, D. (1995). Graphical Modeling. In W. J. Krzanowski (Ed.), *Recent advances in descriptive multivariate analysis* (pp. 135–156). Oxford, New York: Clarendon Press.

Fischer, G., Rohde, M., & Wulf, V. (2007). Community-based learning: The core competency of residential, research-based universities. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 9–40.

Francis, J. G. F. (1961). The QR transformation, I. *The Computer Journal, 4*(3), 265–271.

Glassman, M., & Kang, M. J. (2011). The logic of wikis: The possibilities of the Web 2.0 classroom. *International Journal of Computer-Supported Collaborative Learning, 6*(1), 93–112.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery.* Boston: Kluwer Academic Publishers.

Hung, D., Lim, K. Y. T., Chen, D.-T. V., & Koh, T. S. (2008). Leveraging online communities in fostering adaptive schools. *International Journal of Computer-Supported Collaborative Learning, 3*(4), 373–386.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy.* New York: Wiley.

Kienle, A., & Wessner, M. (2006). An analysis of the CSCL community: Development of participation. *International Journal of Computer-Supported Collaborative Learning, 1*(1), 9–33.

Kim, S.-S., Kwon, S., & Cook, D. (2000). Interactive visualization of hierarchical clusters using MDS and MST. *Metrika, 51*(1), 39–51.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling. Sage University Paper series on Quantitative Application in the Social Sciences, 07-011.* Beverly Hills: Sage Publications.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics, 24*(5), 719–720.

Larusson, J. A., & Altermann, R. (2009). Wikis to support the "collaborative" part of collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 4*(4), 371–402.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management, 41*, 1462–1480.

Lund, A., & Rasmussen, I. (2008). The right tool for the wrong task? Match and mismatch between first and second stimulus in double stimulation. *International Journal of Computer-Supported Collaborative Learning, 3*(4), 387–412.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California Press.

Nett, B. (2008). A community of practice among tutors enabling student participation in a seminar preparation. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 53–67.

Pifarré, M., & Staarman, J. K. (2011). Wiki-supported collaborative learning in primary education: How a dialogic space is created for thinking together. *International Journal of Computer-Supported Collaborative Learning, 6*(2), 187–205.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Price, D., & Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist, 21*, 1011–1018.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York: Mc Graw-Hill, Inc.

Spearman, C. (1950). *Human ability.* London: Macmillan.

Stahl, G. (2002). Contributions to a theoretical framework for CSCL. In G. Stahl (Ed.), *Proceedings International Conference of Computer Supported Collaborative Learning, CSCL'2002* (pp. 62–71).

Steinbach, M., Karypis, G. & Kumar, V. (2000). *A comparison of document clustering techniques*. University of Minnesota, Technical report *#00-034*. http://www.cs.fit.edu/~pkc/classes/ml-internet/papers/ steinbach00tr.pdf

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*, 163–171.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science, 49*, 327–355.

Zellig, H. (1954). Distributional structure. *Word, 10*(2/3), 146–162.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge: Harvard University Press.