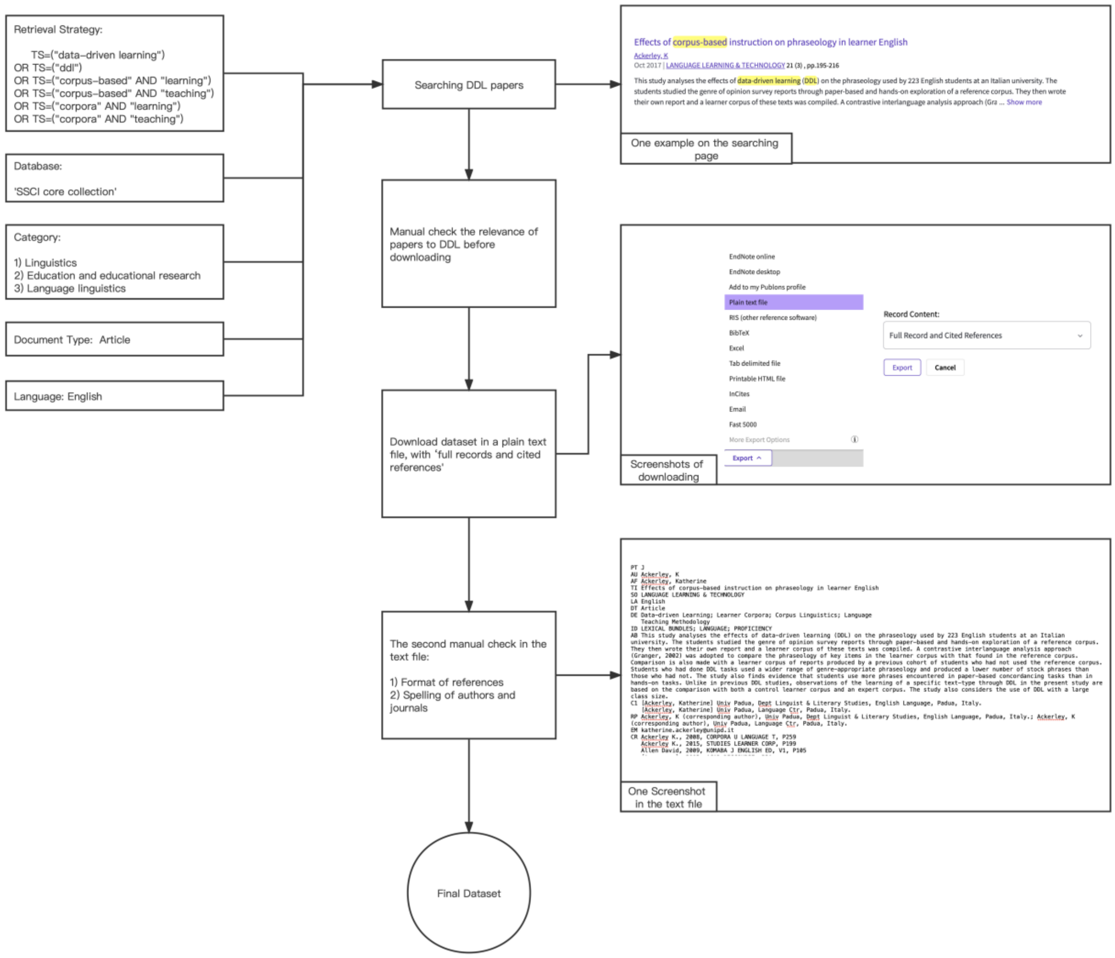
Appendix A
Procedures and parameters used in this study


Retrieval Strategy:
TS=("data-driven learning")
OR TS=("ddl")
OR TS=("corpus-based" AND "learning")
OR TS=("corpus-based" AND "teaching")
OR TS=("corpora" AND "learning")
OR TS=("corpora" AND "teaching")
Database:
'SSCI core collection'
Category:
1) Linguistics
2) Education and educational research
3) Language linguistics
Document Type: Article
Language: English
Searching DDL papers
Manual check the relevance of papers to DDL before downloading
Download dataset in a plain text file, with 'full records and cited references'
The second manual check in the text file:
1) Format of references
2) Spelling of authors and journals
Final Dataset
One example on the searching page
Screenshots of downloading
EndNote online
EndNote desktop
Add to my Publons profile
Plain text file
RIS (other reference software)
BibTeX
Excel
Tab delimited file
Printable HTML file
InCites
Email
Fast 5000
More Export Options
Export
Cancel
One Screenshot in the text file

Effects of corpus-based instruction on phraseology in learner English
Ackerley, K
Oct 2017 | LANGUAGE LEARNING & TECHNOLOGY 21 (3) , pp.195-216
This study analyses the effects of data-driven learning (DDL) on the phraseology used by 223 English students at an Italian university. The students studied the genre of opinion survey reports through paper-based and hands-on exploration of a reference corpus. They then wrote their own report and a learner corpus of these texts was compiled. A contrastive interlanguage analysis approach (Gra ... Show more

Record Content: Full Record and Cited References

```
PT J
AU Ackerley, K
AF Ackerley, Katherine
TI Effects of corpus-based instruction on phraseology in learner English
SO LANGUAGE LEARNING & TECHNOLOGY
LA English
DT Article
DE Data-driven Learning; Learner Corpora; Corpus Linguistics; Language
   Teaching Methodology
ID LEXICAL BUNDLES; LANGUAGE; PROFICIENCY
AB This study analyses the effects of data-driven learning (DDL) on the phraseology used by 223 English students at an Italian
university. The students studied the genre of opinion survey reports through paper-based and hands-on exploration of a reference corpus.
They then wrote their own report and a learner corpus of these texts was compiled. A contrastive interlanguage analysis approach
(Granger, 2002) was adopted to compare the phraseology of key items in the learner corpus with that found in the reference corpus.
Comparison is also made with a learner corpus of reports produced by a previous cohort of students who had not used the reference corpus.
Students who had done DDL tasks used a wider range of genre-appropriate phraseology and produced a lower number of stock phrases than
those who had not. The study also finds evidence that students use more phrases encountered in paper-based concordancing tasks than in
hands-on tasks. Unlike in previous DDL studies, observations of the learning of a specific text-type through DDL in the present study are
based on the comparison with both a control learner corpus and an expert corpus. The study also considers the use of DDL with a large
class size.
C1 [Ackerley, Katherine] Univ Padua, Dept Linguist & Literary Studies, English Language, Padua, Italy.
   [Ackerley, Katherine] Univ Padua, Language Ctr, Padua, Italy.
RP Ackerley, K (corresponding author), Univ Padua, Dept Linguist & Literary Studies, English Language, Padua, Italy.; Ackerley, K
(corresponding author), Univ Padua, Language Ctr, Padua, Italy.
EM katherine.ackerley@unipd.it
CR Ackerley K., 2008, CORPORA U LANGUAGE T, P259
   Ackerley K., 2015, STUDIES LEARNER CORP, P199
   Allen David, 2009, KOMABA J ENGLISH ED, V1, P105
```

Figure A1. The procedures of data collection

Figure A1 depicts the procedures of data collection based on the WoS core collection database. In retrieving the data set, the following steps were taken. Firstly, the function 'advanced search' was used to implement the following retrieval strategy:

> TS=("data-driven learning") OR TS=("ddl") OR TS=("corpus-based" AND "learning") OR TS=("corpus-based" AND "teaching") OR TS=("corpora" AND "learning") OR TS=("corpora" AND "teaching")

The results show that the earliest paper in this field is from 1994 and the most recent paper is from 2021 (retrieved in Nov 2021). Secondly, in order to refine the results to research articles related to the field of applied linguistics, three categories were used to filter the dataset: 1) Linguistics; 2) Education and educational research; 3) Language linguistics; and alternative publication types (e.g., dissertations and theses, book reviews, and editorials) were removed by selecting document type of 'articles' following Liu and Hu (2021).

Thirdly, a close check of the titles, keywords and abstracts of the articles was conducted to confirm their close relevance to DDL, and articles irrelevant to DDL were excluded. The initial search identified 412 articles, which were then subjected to a manual check and excluded the articles that were not relevant to the topics. For example, one of the searched results was *Machine learning comprehension grammars for ten languages* authored by Suppes, Bottner and Liang (1996). Although it contains 'corpora' and 'learning' in its *abstract* and *title* respectively, it has no relevance to DDL, and thus was excluded. This is then followed by ticking the articles that are relevant to the topics. One example relevant to DDL in our dataset was Ackerley (2017), which is highly related to DDL based on close examination, and its screenshot in WoS is presented in Figure A1. Finally, a total of 126 articles were selected with 3297 distinct references.

The next step is downloading papers as a plain text file from WoS, with 'full records and cited references' (see Figure A1). Then, for the maximum utility and consistency of the dataset, formatting errors and misspellings were also corrected in the text file. For instance, author names with full letters capitalised were changed into the first letter capitalised, such as 'BOULTON, A' to 'Boulton, A'.

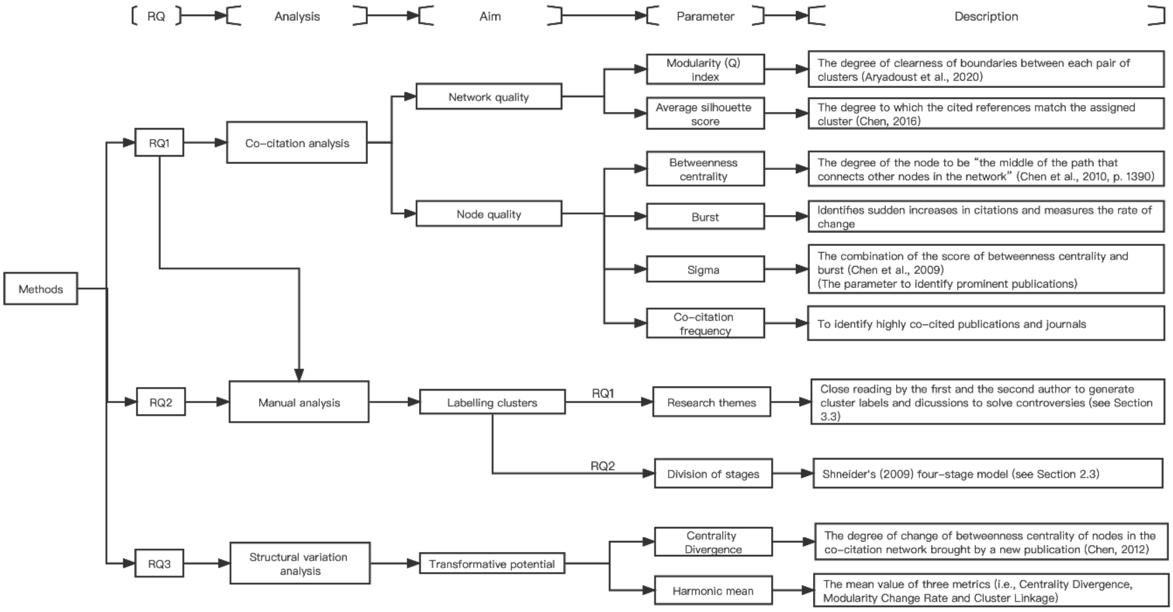

RQ
Analysis
Aim
Parameter
Description
Methods
RQ1
Co-citation analysis
Network quality
Modularity (Q) index
The degree of clearness of boundaries between each pair of clusters (Aryadoust et al., 2020)
Average silhouette score
The degree to which the cited references match the assigned cluster (Chen, 2016)
Node quality
Betweenness centrality
The degree of the node to be "the middle of the path that connects other nodes in the network" (Chen et al., 2010, p. 1390)
Burst
Identifies sudden increases in citations and measures the rate of change
Sigma
The combination of the score of betweenness centrality and burst (Chen et al., 2009) (The parameter to identify prominent publications)
Co-citation frequency
To identify highly co-cited publications and journals
RQ2
Manual analysis
Labelling clusters
RQ1
Research themes
Close reading by the first and the second author to generate cluster labels and dicussions to solve controversies (see Section 3.3)
RQ2
Division of stages
Shneider's (2009) four-stage model (see Section 2.3)
RQ3
Structural variation analysis
Transformative potential
Centrality Divergence
The degree of change of betweenness centrality of nodes in the co-citation network brought by a new publication (Chen, 2012)
Harmonic mean
The mean value of three metrics (i.e., Centrality Divergence, Modularity Change Rate and Cluster Linkage)

Figure A2. Parameters used in the present study

The description of parameters used in addressing each question is presented below.

CiteSpace (6.1.R2) (Chen, 2006) was used in the co-citation analysis and SVA. The Log-Likelihood Ratio (LLR) algorithm was utilised to identify the characteristics of clusters, as previous studies had confirmed its utility (Chen et al., 2010). In addition, following Chen (2013), the 'keywords' function (i.e., keywords selected by the authors and WoS indexing terms) was used to determine automatic cluster labels.

To answer the first research question, both the network structure and nodes (i.e., cited publications in the network) were analysed. First, following Chen et al. (2010), the modularity (Q) index and average silhouette score were adopted to measure the quality of the network, which are automatically computed in CiteSpace. The modularity scores (0 to 1) determine the clearness of boundaries between each pair of clusters, which refers to the extent to which the network can be decomposed into different components (Aryadoust et al., 2020). The network can be decomposed into several recognisable clusters if the modularity score of the network is close to 1. The average silhouette value (-1 to 1) determines the quality of a clustering structure, which refers to the degree to which the cited references match the assigned cluster (Chen, 2016). A high average silhouette score indicates the high reliability of clustering. The modularity score of the network in this study was 0.79, indicating clear boundaries between each pair of clusters; and the high quality of clustering configuration is attested by the weighted mean silhouette score of 0.92, indicating that the cited publications in the co-citation network can be clearly separated into different clusters (see Figure A3).

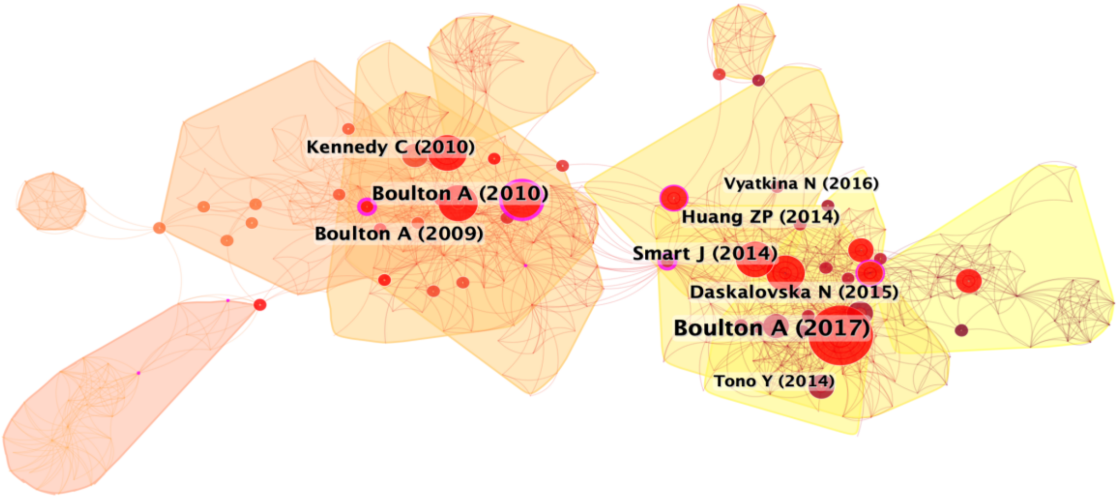


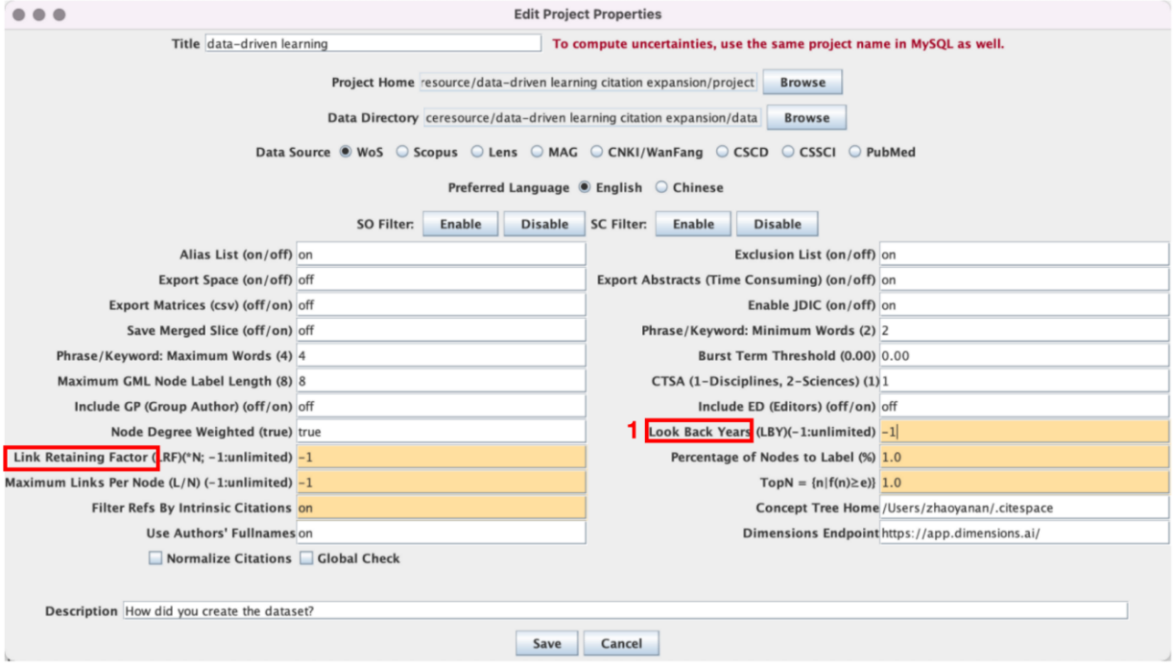
Figure A3. The co-citation network generated in CiteSpace

To identify important research themes and prominent publications, the following three metrics were enabled in CiteSpace: sigma (∑), betweenness centrality and burst. Betweenness centrality (0 to 1) specifies the degree to which the node publication is "in the middle of the path that connects other nodes in the network" (Chen et al., 2010, p. 1390). A node of high betweenness centrality indicates a strong connection between two or more large clusters of nodes, with this node in-between. Burst is an indicator of recent or ongoing research interest according to sudden increases in citations and measures the rate of change. A large number of bursts in one cluster indicate a high level of activity in the respective research area (Chen, 2016). Sigma, a measure by combining the value of betweenness centrality and citation burst (Chen, Chen, Horowitz, Hou, Liu & Pellegrino, 2009), was used to identify prominent publications of high innovative potential. A high value of sigma indicates a high level of scientific novelty (Chen et al., 2009), and thus the more prominent the publication. The nodes with high centrality and burst are signalled in purple and red respectively, as can be seen in Figure A3. For example, the node representing Boulton and Cobb (2017) is red and is the largest node, indicating its burst status and high co-citation frequency; while Boulton (2010a) obtained a high centrality score, which is indicated by its purple colour. Also of note, although several publications were not identified as prominent publications by automatic analysis in CiteSpace, they obtained relatively high co-citation frequencies, indicating their high impact. The inclusion of frequently co-cited papers enabled a wider coverage of high-impact publications in the DDL field.

In answering the second research question, based on the existing themes of publications in each cluster, the major clusters were assigned to appropriate stages of Shneider's (2009) evolutionary model in terms of their time frames, interrelationships, and embodiment of the model's defined qualities for each stage. For instance, if the manual analysis reported that a specific cluster included several publications focusing on the research instrument, such as the invention of the software and corpus, then the cluster can be regarded as representing the second stage. However, if the research instrument(s) is utilised in the publications to investigate new subject domains, it can be deduced that the cluster has reached the third stage.

To answer the third question, SVA was employed to detect transformative publications in co-citation networks by means of Centrality Divergence (C_{KL}) and Harmonic mean (H) scores. C_{KL} is a parameter that measures the degree of change of betweenness centrality of nodes in the co-citation network brought by a new publication (Chen, 2012). C_{KL} was adopted as the main parameter as it has been identified as optimal in detecting the transformative characteristics of a particular publication at cross-disciplinary levels (Sebastian & Chen, 2021). H is a parameter that signals the potential impact of a publication from various aspects of structural variations. Thus, it can be regarded as a complementary metric to C_{KL} as it can detect papers that are outstanding concerning their average scores of different metrics.

Appendix B
The detailed setting and operating of CiteSpace


Edit Project Properties
To compute uncertainties, use the same project name in MySQL as well.
Browse
Browse
SO Filter: Enable Disable
SC Filter: Enable Disable
1
2
Save
Cancel

Title: data-driven learning

Project Home: resource/data-driven learning citation expansion/project

Data Directory: ceresource/data-driven learning citation expansion/data

Data Source: ● WoS ○ Scopus ○ Lens ○ MAG ○ CNKI/WanFang ○ CSCD ○ CSSCI ○ PubMed

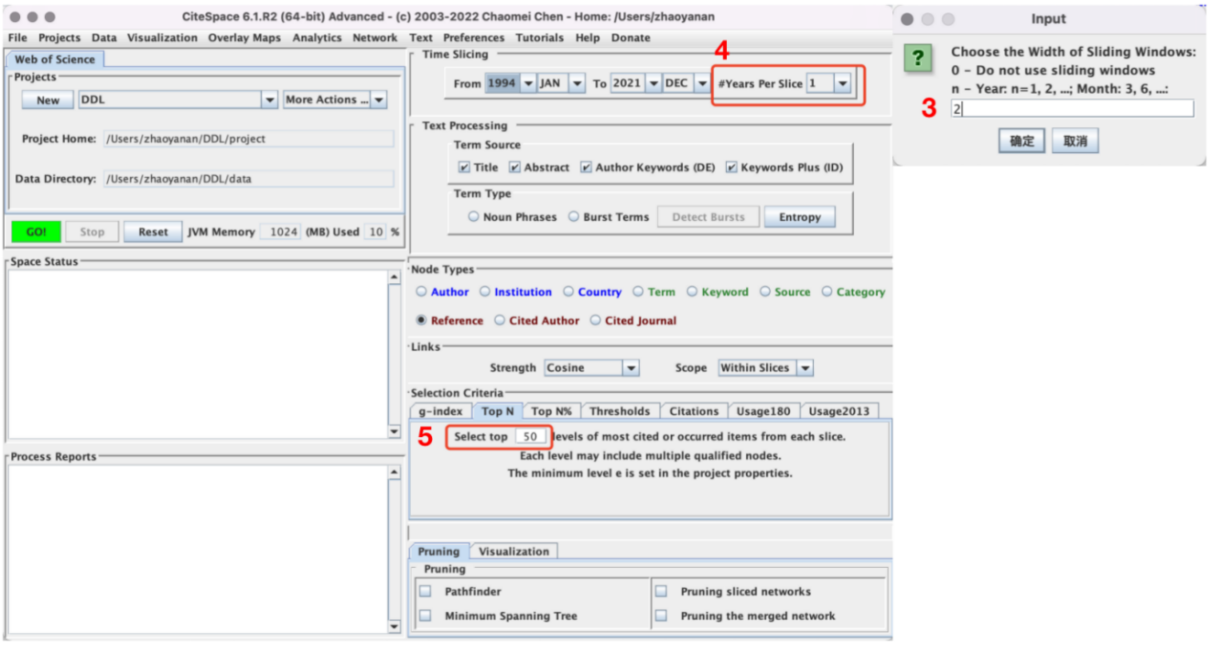
Preferred Language: ● English ○ Chinese

Alias List (on/off)	on	Exclusion List (on/off)	on
Export Space (on/off)	off	Export Abstracts (Time Consuming) (on/off)	on
Export Matrices (csv) (off/on)	off	Enable JDIC (on/off)	on
Save Merged Slice (off/on)	off	Phrase/Keyword: Minimum Words (2)	2
Phrase/Keyword: Maximum Words (4)	4	Burst Term Threshold (0.00)	0.00
Maximum GML Node Label Length (8)	8	CTSA (1-Disciplines, 2-Sciences) (1)	1
Include GP (Group Author) (off/on)	off	Include ED (Editors) (off/on)	off
Node Degree Weighted (true)	true	Look Back Years (LBY)(-1:unlimited)	-1
Link Retaining Factor (LRF)(*N; -1:unlimited)	-1	Percentage of Nodes to Label (%)	1.0
Maximum Links Per Node (L/N) (-1:unlimited)	-1	TopN = {n|f(n)≥e)}	1.0
Filter Refs By Intrinsic Citations	on	Concept Tree Home	/Users/zhaoyanan/.citespace
Use Authors' Fullnames	on	Dimensions Endpoint	https://app.dimensions.ai/

□ Normalize Citations □ Global Check

Description: How did you create the dataset?

Figure B1. Properties set in CiteSpace



CiteSpace 6.1.R2 (64-bit) Advanced - (c) 2003-2022 Chaomei Chen - Home: /Users/zhaoyanan

Projects: DDL

Project Home: /Users/zhaoyanan/DDL/project

Data Directory: /Users/zhaoyanan/DDL/data

JVM Memory 1024 (MB) Used 10 %

Time Slicing: From 1994 JAN To 2021 DEC #Years Per Slice 1

Term Source: ☑ Title ☑ Abstract ☑ Author Keywords (DE) ☑ Keywords Plus (ID)

Term Type: ○ Noun Phrases ○ Burst Terms

Node Types: ○ Author ○ Institution ○ Country ○ Term ○ Keyword ○ Source ○ Category ● Reference ○ Cited Author ○ Cited Journal

Links: Strength Cosine Scope Within Slices

Selection Criteria: Top N

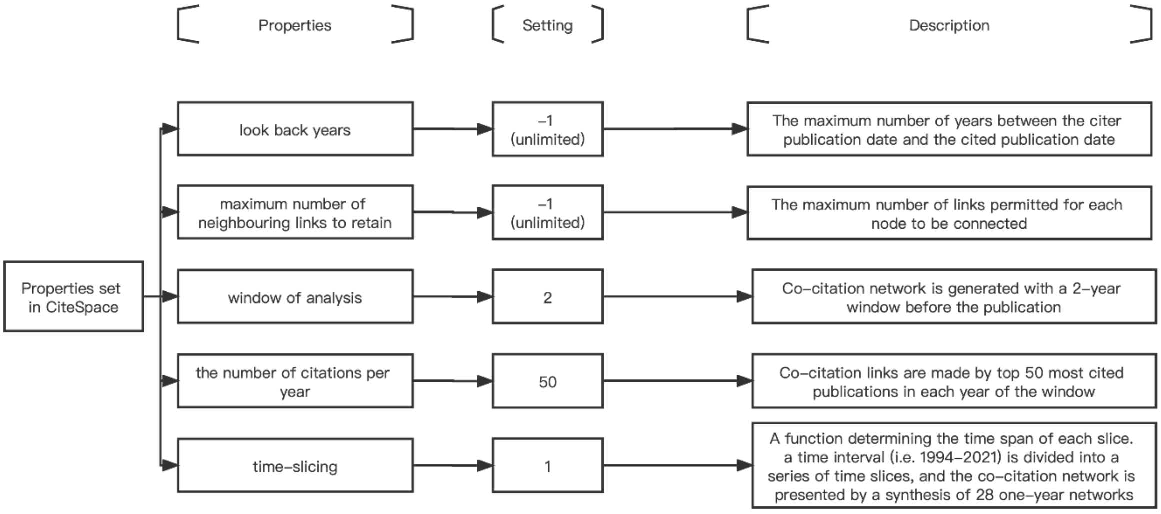
Select top 50 levels of most cited or occurred items from each slice.
Each level may include multiple qualified nodes.
The minimum level e is set in the project properties.

Pruning: □ Pathfinder □ Minimum Spanning Tree □ Pruning sliced networks □ Pruning the merged network

Input

Choose the Width of Sliding Windows:
0 – Do not use sliding windows
n – Year: n=1, 2, ...; Month: 3, 6, ...:
2

Figure B2. The main user interface of CiteSpace



Properties	Setting	Description
look back years	-1 (unlimited)	The maximum number of years between the citer publication date and the cited publication date
maximum number of neighbouring links to retain	-1 (unlimited)	The maximum number of links permitted for each node to be connected
window of analysis	2	Co-citation network is generated with a 2-year window before the publication
the number of citations per year	50	Co-citation links are made by top 50 most cited publications in each year of the window
time-slicing	1	A function determining the time span of each slice. a time interval (i.e. 1994-2021) is divided into a series of time slices, and the co-citation network is presented by a synthesis of 28 one-year networks

Figure B3. Properties set in the present study

Figure B1-B3 illustrate the properties set in CiteSpace and the explanation of these properties. In Figure B1, the first property, ‘look back years’, refers to the maximum number of years between the citer publication date and the cited publication date. Setting the value as ‘-1’ stands for the unlimited number of years, ensuring wider coverage of the synthesized network. The second property, ‘link retaining factor’, is the maximum number of links permitted for each node to be connected. Similarly, setting this property to ‘-1’ means that links connected between each pair of nodes were unlimited, reaching a more complete picture of groups of nodes (Chen, 2016). The third, fourth and fifth properties were presented in Figure B2. The third property, ‘the window of the analysis’, refers to the window considered from the publishing year. Following Chen, Hu, Liu and Tseng (2012), this property was set to 2, indicating that the two years [Y-2, Y-1] before the publishing year Y of the article were used for the network generation. The fourth property, top N is the number of most-cited articles in each year to be used to generate the network. Setting this value as default ‘50’ means that the 50 most cited publications in each year were used to construct the co-citation network. The fifth property, the ‘time-slicing’ function is a function determining the time span of each slice. Setting the value to ‘1’ stands that a time interval (i.e., 1994-2021) is divided into a series of time slices, and the co-citation network is presented by a synthesis of 28 one-year networks (1994-2021) (Chen, 2017). Thus, the co-citation network was constructed based on “multiple panoramic snapshots” (Chen, 2016, p. 49). A detailed description of the properties and the value set are presented in Figure B3.

Appendix C

Figures for Co-cited publications and Co-citation network of journals in results

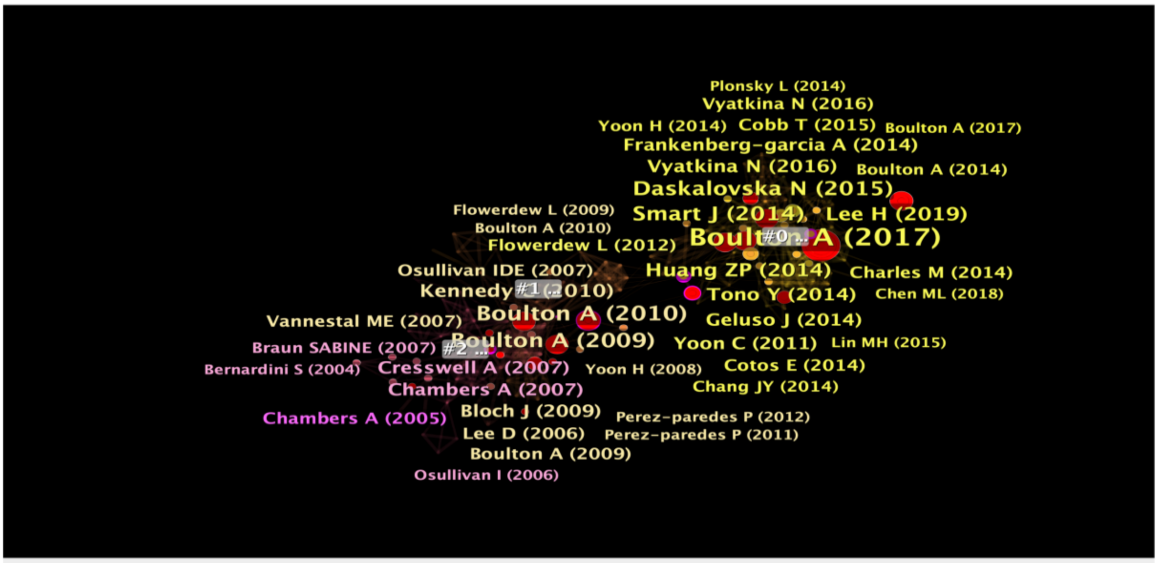


Figure C1. Co-cited publications in Cluster #0, #1 and #2

Figure C1 presents co-cited publications in the three biggest clusters: #0, #1 and #2. The font size of each publication indicates their co-citation frequency. In other words, the larger the font size, the more frequently co-cited the publication is. For instance, Boulton and Cobb (2017) is the most frequently co-cited publication in Cluster #0 and is illustrated in the largest font size.

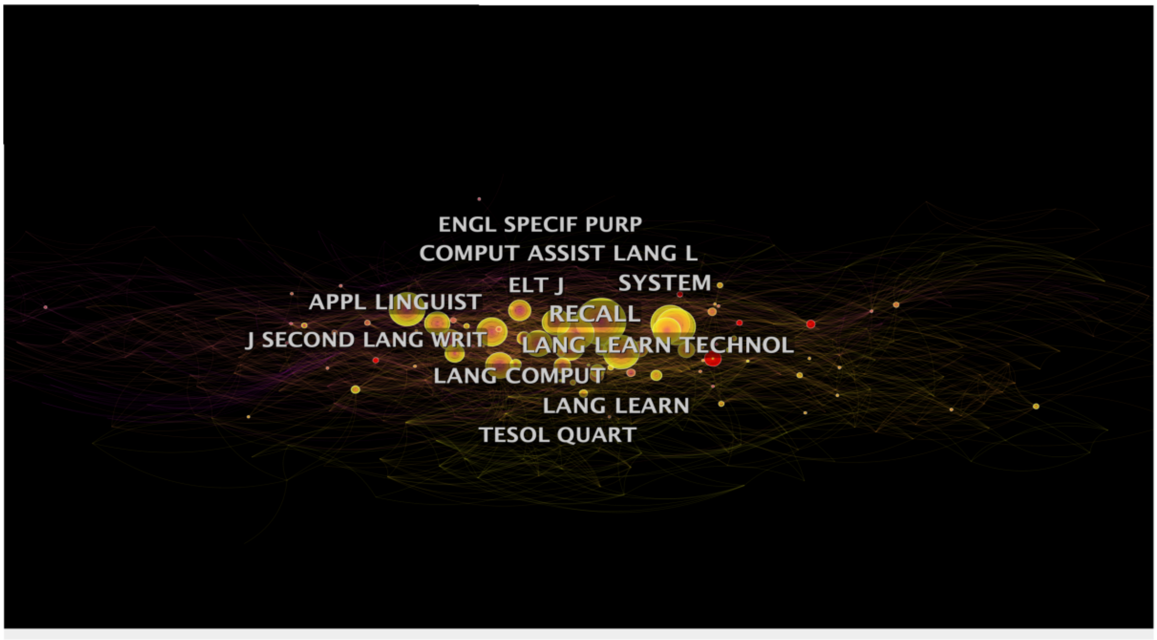


Figure C2. Co-citation network of journals

Figure C2 is the screenshot of highly co-cited journals or books, and this is addressed in detail in Section 4.4.

References for appendices

Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology, 21*(3), 195–216.

Chen, C. (2016). *CiteSpace: A practical guide for mapping scientific literature.* Nova Science Publishers, Inc.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191–209. https://doi.org/10.1016/j.joi.2009.03.004

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy, 12*(5), 593–608. https://doi.org/10.1517/14712598.2012.674507

Suppes, P., Bottner, M., & Liang, L. (1996). Machine Learning Comprehension Grammars for Ten Languages. *Computational Linguistics, 22*(3), 329–350.