EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

EUROPEAN NETWORK OF e-LEXICOGRAPHY

COST ENeL
Working Group 4 Task Group „Meta-Lexicography”

General information

Task Group Name
Meta-Lexicography

Main Working Group
WG4

Liaison with WG(s) / task groups

WG 1

The Task Group needs to keep in touch with the WG 1 in order to feed discussion on the ENeL Dictionary Portal (henceforth, EDP). WG 1 experts on dictionary use may help to structure work on the desired forms of dictionary access, optimal search strategies and effective ways of improving user’s experience.

WG 3

The Task Group should collaborate with the WG 3 in order to identify the cutting-edge projects in the field of electronic lexicography, the advantages of incorporating corpus data into the integrated dictionary, and possible contribution of computational linguistics.

WG 4 Task Groups

I. Task Group 1. Vocabulary of Emotions | Task Group 2. Lexical Variation | Task Group 3. Pan-European Vocabulary | Task Group 7. Dictionary of European Concepts | Task Group 8. Common European Heritage of Vocabularies / Etymology

Collaboration should help to identify:

1) main drawbacks and ways of improving on current lexicographic practice to better meet needs of researchers on language contact and change;
2) promising representation techniques that may adequately reflect or reveal language contact and change patterns;
3) optimal dictionary access methods;
4) possible entities of data integration (etymology, reconstructed form, concept etc.).

II. Task Group 4. Digital Humanities

Collaboration should help to identify:

1) innovative ways of representing dictionary content;
2) specific needs of the digital humanities community.

III. Task Group 9. Liaison with *Atlas Linguarum Europae*

Collaboration may help to identify suggested means of representing geographical dimension of the Pan-European vocabulary.

IV. Task Group 12. Metadata, eInfrastructures, Standards

Collaboration should help to identify:

1) formats of data exchange which are best adapted to meet the Task Group's objectives;
2) drawbacks and advantages of emerging and long-established standards of dictionary content description, with special regard to its spatio-temporal dimension;
3) best practices and cutting-edge examples of non-lexicographic data exchange initiatives.

V. Task Group 13. European Languages Portal

Collaboration may help to exchange experience in building integration tools.

Coordinator
Krzysztof NOWAK

Proponents
Krzysztof NOWAK

Participants
As of 20 January 2015 the Task Group consists of 3 persons:

- Bruno BON,
- Nathalie MEDERAKE,
- Krzysztof NOWAK.

WG 4 members who wish to work within the framework described below are very welcome to join us!

Content

Current discussions on lexicographic data integration often focus on the technical side of the enterprise, with special emphasis being on exchange formats or encoding schemes. As for emerging data linking projects, rather than aiming at human users, they usually need to be embedded in larger text processing infrastructures. This is where electronic lexicography may come in, providing both expert and general public with meaningful and structured insights into still growing **quantity of lexicographic data**. The Task Group on Meta-Lexicography addresses both theoretical and practical challenges that integrated retrieval of heterogeneous dictionary content may pose, whether it be within the EDP or outside it.

As for now following issues have been identified and will be addressed by the Task Groups Members:

1. Dictionary content access

A convenient point of departure might be existing research on single electronic dictionary use which, if extrapolated, should help to understand what information may the users of the EDP be looking for and what are the major design decisions to be taken to make the EDP a user-friendly tool. Obviously, otherwise than in the single-language dictionaries, the EDP will have to handle different European scripts, varying lexicographic practices etc. As such it seems reasonable to call for adopting data-oriented approach and heavy processing of the **user input** (e.g. language recognition, spell-checking, search suggestions etc.).

Yet, if the EDP has to go beyond mere aggregation of dictionary content, the problem might arise of what is its basic entity of data linking and retrieval. Taking hypothetical or dead-language etyma as a basis, albeit self-evident at a first glance, not only may require substantial manual alignment, but also is of limited use as far as non-expert user access and interface is concerned. Whatever the finally adopted solution will be, in order to be stored and efficiently queried, those entities must be attributed **unique identifiers**. Since the EDP does not exclude widely used historical languages (such as Latin) and since scholarly dictionaries of modern languages are rich in **spatio-temporal information**, this dimension also should be properly reflected in the integration model, whether it be for monitoring language change or enabling investigation into historical background of the current lexical patterns.

All in all, it is highly desirable to seek, test and report on the long-established and recently emerged **non-lemmatic dictionary access** forms, such as thesauri, ontology- or Wordnet-aided conceptual search, map browsing, pattern look up etc. Other easily retrievable candidates for integrated access might be also source quotations, historical periods, language variation patterns, and so on. Since the project should attract varied audience, several well-designed knowledge retrieval **paths** and attractive **narratives** need to be offered to the users, depending on their goals and search strategies.

2. Research-driven integration

By no means is integration of lexicographic exclusively a matter of exchange formats or efficient data storage. A means rather than an end, it contributes to emergence of new research questions and reshapes daily practice of lexicographers.

a. The general user convenience and needs being an important concern, the integration of lexicographic data should be, in the same time, research-driven. It means, firstly, that it needs to be guided by **precise research questions** and to meet research community expectations as well. Work should be done in order to demonstrate which of the currently discussed research problems may

benefit from dictionaries' integration and how they are addressed by the relevant EDP's design choices. Secondly, the Task Group Members will argue that the integrated lexicographic resources point research community towards **new directions of scholarly investigation** and may cause that new research questions to emerge. Thirdly, further research needs to be carried out in order to answer: 1) whether the data integration does not happen to create new research objects; 2) what are the consequences of lexicographic data being often scarce and incomplete; 3) what is the precise linguistic reality that the integrated data account for in a coherent manner. Finally, the Task Group Members will discuss whether data-oriented integration of the dictionary content is not a step towards some „theory-agnostic" lexicography which would allow for data reuse within changing theoretical frameworks.

b. Despite not being 'big data' in the strict sense of the term, integrated lexicographic data will probably not lend themselves to more traditional, word-by-word inspection any more. It means that both general public and scholarly community needs to be equipped with tools of **exploratory analysis**, such as summaries, visualisations, tables, timelines, which should be considered as research tool on their own rather than decorative addition. As such, not only should they help to reduce the quantity of lexicographic data, but also render otherwise obscure (chronological, geographical etc.) **patterns** of language change (lexical loans and lacunas, varying orthography etc.) visible.

c. Data integration will eventually lead to changing the current landscape of scholarly lexicography, and that for two reasons. On the one hand, when taking into account linked data, lexicographers will gain an invaluable help in dictionary writing process. Thanks to the more **convenient access** to linguistic information, better insight into semantic change patterns etc., the process of writing dictionary entry should be facilitated and refined, even at the expense of time spent on analysing newly acquired evidence. On the other hand, data integration initiatives may lead electronic lexicographers to enhance their **data encoding schemes**, to neglect information that can be easily inferred from other sources.

3. Emerging tools and examples to follow

In the last years, a substantial body of work has been done on data exchange and standardisation (both being a topic of a separate WG4 Task Group). Moreover, whether it be in XML, JSON, or RDF form, data can be nowadays stored in a variety of database tools. As far as their display is concerned, with the advent of easy to implement JS libraries (such as d3.js and similar), GiS modelling tools, or robust web presentation platforms, researchers and lay users may be offered interactive insights into linguistic data. Emphasis being on **free and open-source solutions**, selected types of existing storage and presentational tools (e.g. wiki, no-SQL database etc.) need to be tested to examine their compatibility with integrated dictionary data model. Basing on use scenarios further guidelines for the EDP design may follow.

4. Encyclopaedia-dictionary interface

Integration will certainly lead to fuller exploitation of the **potential which remains hidden** in the dictionaries' entries. Yet, effort should be made to seek opportunities for further enhancements of the EDP with external resources. One of the most promising directions seems to be enriching original data by means of encyclopaedias, databases and other **knowledge resources**. The encyclopaedic data may contribute to better discourse comprehension and production, since, as many studies show, linguistic and world knowledge tend to interweave in the human search for meaning. As far as research community is concerned, adopting such an approach allows for parallel investigation of language patterns and **social, institutional, historical change**.

Scope
(Where and - if so - how does your task group meet the objectives needs?)

1. WG4 Objective: *Developing editorial guidelines for the integration of European information into more traditional and into innovative e-dictionaries*

Research described above in section Content §1, §2a,c and §3 meets the present objective.

2. WG4 Objective: *Developing ways in which already existing information from single language dictionaries can be displayed and interlinked to represent more adequately their common European heritage*

Research described above in section Content §1 and §2b meets the present objective.

3. WG4 Objective: *Finding new applications for the very large amount of interconnected dictionary information from the European dictionary portal in the field of digital humanities*

Research described above in section Content §2a and §4 meets the present objective.

General contribution to COST ENeL
1. Memorandum of Understanding: *Establishing new ways of representing the common heritage of the languages of Europe. Giving users easier access to scholarly dictionaries and to bridge the gap between the general public and scholarly dictionaries.*

and

3. Memorandum of Understanding: *Developing a common approach to e-lexicography that forms the basis for a new type of lexicography that fully embraces the pan-European nature of much of the vocabularies of the languages spoken in Europe.*

Work described above in section Content §1 and §2 aims at identifying needs of both research and general public. Through case studies, mashups and prototypes Members of the Task Group will demonstrate how different electronic frameworks, such as wiki, and visualisation tools may bring into relief the common heritage and the variety of European languages in its synchronic and diachronic dimension.

2. MoU: *Establishing both a broader and more systematic exchange of expertise and common standards and solutions.*

While searching for meaningful non-textual representation of lexicographic knowledge, the Task Group Members will attempt to suggest possible ways of collaboration between lexicographers, digital humanists and computational linguists. They will also argue for closer collaboration with data specialists and for community-driven standardisation of lexicographic content description.

4. MoU: *Frame a network: Connect with Stakeholders, relevant other projects etc.*

Members of the Task Group will attempt to identify possible sources of collaborative funding. Originating from the historical lexicography community they attempt to reach broader public and argue for inclusion of innovative historical dictionaries into the EDP.

Relation to COST ENeL WG4 objectives
(in detail: does your task group meet this needs and - if so - how. Please, give examples concerning your proposed outcome and suggest deliverables e.g. guidelines, report, scientific article, handbook)

→ The Task Group „Meta-Lexicography", despite of addressing a number of theoretical questions, is practice-based and employs mainly inductive methods. Case studies, online showcases, working prototypes, mashups etc. will form, at the same time, output and point of departure for further discussion. The Task Group will also communicate with other Task Groups and WGs to get fuller insight into theoretical and practical issues concerning data integration.

1. *Eurolinguistics: Studying the migration and re-migration of words and meanings across the languages of Europe*

→ Depending on the empirical context each Task Group member is working within, language contact and change patters will be demonstrated in **reports** or **case studies** or **working prototypes** illustrating European lexical loans (N. Mederake) and development of the Medieval Latin vocabulary (B. Bon, K. Nowak).
More: Research described above in section Content §2b meets the present objective.

2. *Explore the possibilities of extensive interlinking of dictionary content from different European languages*

→ B. Bon and K. Nowak will present **case study** of linking between several European dictionaries of Medieval Latin. Problems which arise and perspectives of further development may be, if needed, put into short **report**.
More: Research described above in section Content §1 meets the present objective.

3. *Consider how extensive interlinking can generate new lines of research in the field of digital humanities*

→ N. Mederake will test opportunities that wiki-based tools offer in asking new questions about language contact. K. Nowak will demonstrate how alternative ways of lexicographic content representation may contribute to our knowledge of language change. Short **case studies** or **mashups** may be used to present the output.
More: Research described above in section Content §2a meets the present objective.

4. *Develop shared editorial practices among the dictionaries of Europe*

→ Research described above in sections Content §2a-c partially meets the present objective. Short **case studies** or **mashups** may be used to present the output. If mature enough, they can be converted into more extensive **guidelines**. Possible overlap, however, with the Task Group on e-Infrastructures should be avoided, with the Task Group „Meta-Lexicography" representing point of view of practising lexicographers.

5. *Develop a roadmap to possible ways for the extensive linking and interconnection of the data in European dictionaries in the European dictionary portal that will generate new lines of research in the field of digital humanities*

→ Research described above in section Content §2a-c meets partially the objective. Possible overlap with the Task Group on „Digital Humanities" should be avoided, with the present Group providing insights without referring to any precise scenario.

6. *Provide a foundation for the further exploration of a Pan-European approach to lexicography by discussing new standards and methodologies to describe the common European heritage (how to provide a pan-European view on the vocabularies of the languages of Europe?)*

→ Research described above in section Content §4 meets the present objective.

7. *Publish scholarly articles in joint teams.*

→ Opportunities for common publication are still to emerge.

8. *Promote sustainability by identifying potential funding sources and developing collaborative funding applications.*

→ Collaborative funding opportunities are being actively sought. Yet, general ENeL guidelines concerning open project calls would be very welcome.

9. *Does your task group take into consideration gender issues?*

→ As a research problem: no, at least not for the moment. As far as gender balance is concerned, the Task Group currently consists of 2 Males and 1 Female. General ENeL guide would be welcome in order to better address gender issues.

10. *Does your task group meet the needs on visually impaired people?*

→ Not for the moment. ENeL guidelines concerning visually impaired people's needs would be welcome.

11. *Does your task group take into consideration the general public as users?*

→ Task Group „Meta-Lexicography" examines, among others, efficient and user-oriented ways of dictionary content access. Although a good deal of its work is addressed to scholarly users, it also aims at providing general public with meaningful information, testing new knowledge presentation opportunities and visualisation tools.
More: Research described above in section Content §1 meets the present objective.

12. *Are you open to disseminate your Task Group's results / work in the framework of European children's universities (http://eucu.net/)?*

→ Possibly, more information would be welcome.

Proposal

Partnerships
(reflecting capacities brought by the participants)

- Académie des inscriptions et belles-lettres
- Academy of Sciences in Göttingen
- CNRS
- Polish Academy of Sciences
- Union Académique Internationale

Objectives
(numbered list)

1. Content Access: To suggest alternative methods of dictionary content access in the EDP.
2. Integration Model: To get feedback from other WG4 members concerning desired integration model.
3. Integration Model : To contribute to the discussion about theoretical assumptions and practical consequences of the dictionary content integration.
4. Research-driven Integration: To associate selected research problems with: a) existing and possible visualisation and exploratory analysis tools; b) respective features of the EDP' design (e.g. lexical loan → map etc.).
5. Research-driven Integration: To argue in favour of electronic lexicography as a source of new research questions.
6. Non-lexicographic Data: To demonstrate advantages and drawbacks of encyclopaedia, corpus, database content integration.

Participants

Researcher	Assigned objectives (resp. number)	Empirical context
Bruno **BON**	1, 3, 6	Medieval Latin, European Dictionary of Medieval Latin, WikiLexicographica
Nathalie **MEDERAKE**	1, 3, 4, 5	historical dictionaries entries in consideration of European lexical loans: German, English, Dutch; wikis
Krzysztof **NOWAK**	1, 2, 3, 4, 5, 6	Classical and Medieval Latin, Polish Dictionary of Medieval Latin, WikiLexicographica, Polish Corpus of Medieval Latin, historical lexicography

Activities

Short Term Scientific Missions

- September 2015: Nathalie Mederake → Krzysztof Nowak
- 2016: Krzysztof Nowak →
- 2017: Bruno Bon →

Training Schools

- 2017: WG4 Training School on Linked Data in e-Lexicography and Linguistic Research

Special ideas to support ESR and female researchers

Under discussion.

Workshops, other events

1) Are you planning to connect an event to COST ENeL? If so, which?
    - 2017: if possible, session on the e-lexicography during the International Congress of Medieval Latin (K. Nowak, B. Bon)

2) Are you open to host a COST ENeL event?

    Krzysztof NOWAK, Bruno BON: YES, for example:
    - Workshop on Alternative Forms of Dictionary Content Access
    - Workshop on Dictionary-Encyclopaedia Interface

Other Activities

Under discussion.

Deliverables

2016

- Prototype: Lexicographic Wiki: Linking Dictionaries and Knowledge Resources
- Showcase Report: Alternative representation of dictionary information

2017

- Showcase Report: Non-lemmatic access to the lexicographic content
- Case Studies
    - Language Change and Contact Visualisation

Agenda

2015

- August: working session during the ENeL Sussex meeting
- September: N. Mederake's STSM in Institute of Polish Language, Kraków, Poland

2016

Under discussion.

2017

Under discussion.

Provisions for sustainability

In-kind prevision(s)

(Is, and if so how is the Task Group embedded into institutional or organisational background, e.g. ongoing projects, collaborations, research infrastructures)?

1. All researchers involved in the Task Group activities are employees of public research institutions. Except for the collaboration, their research is financially supported by:

- B. Bon: IRHT-CNRS, France

- N. Mederake: Academy of Science in Göttingen, Germany
- K. Nowak: Institute of Polish Language, Polish Academy of Sciences, Poland

2. All researchers involved in the Task Group are active lexicographers in long-established dictionary enterprises:
- B. Bon: *Novum Glossarium Mediae Latinitatis* (= Dictionary of European Medieval Latin)
- N. Mederake: *Deutsches Wörterbuch's* (revised edition)
- K. Nowak: *(e-)Lexicon Mediae et Infimae Latinitatis Polonorum* (= (Electronic) Dictionary of Polish Medieval Latin).

As such, they participate in respective European networks:
- B. Bon, K. Nowak: *Union Academique Internationale* (Brussels).

3. B. Bon and K. Nowak's collaboration on the *WikiLexicographica* was financed thanks to:
- the ANR Projet „OMNIA. Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins";
- the grants of the Polish Ministry of Science and Higher Education;
- COST Action 1005 „Medioevo Europeo".

Additional funding
(if applicable; are there projects proposed with extra funding to further the results of this task group?)

1. Some parts of the WikiLexicographica infrastructure and respective Medieval Latin dictionaries are being submitted to the DARIAH-PL/DARIAH-FR evaluation.
2. Some parts of the research carried out within the COST ENeL Action will be proposed to:
- the national research agencies,
- the Horizon 2020.

Support

Special needs
Under discussion.

People
Under discussion.

Additional

Profiles of the Task Group Members
1.
- Name: Bruno Bon
- ESR: YES
- Fields of Interest (5 keywords): historical semantics, Medieval Latin
- Fields of Interest (up to 100 words): Bruno Bon is chief-redactor of the *Novum Glossarium Mediae Latinitatis* (UAI – Comité Du Cange). He obtained his PhD from the EPHE (*Les Sermons d'Adémar de Chabannes – édition du manuscrit de Berlin*). Vice-director of the IRHT in charge of digital humanities, he has been coordinating the ANR – Omnia project

(http://glossaria.eu) which resulted in retro-digitisation of several dictionaries (Du Cange's *Glossarium* and the *NGML*) and development of the Treetagger parameters for Medieval Latin lemmatisation. With Krzysztof Nowak, he has been carrying out the semantic *WikiLexicographica* project. He is currently working on: 1) tools for historical semantics and corpus statistics; 2) developing new models of historical lexicography.

- Are you open to further your work in the framework of DARIAH lexical resources working group? YES
- VIAF: http://viaf.org/viaf/90783553

2.

- Name: Nathalie Mederake
- ESR: YES
- Fields of Interest (5 keywords): metalexicography, entry structure, linking, alignment with regard to European lexicography
- Fields of Interest (up to 100 words): Nathalie Mederake is editor and research associate at the *Deutsches Wörterbuch*'s revised edition, Academy of Science in Göttingen. Her interest lies in the meta-structure of historical dictionaries and their possible digital enhancements. With regard to the idea of European loanword lexicography she also focuses on the question how to document the European languages in a way that shows their connections both on a linguistic and a cultural level. Other research interests stem from her PhD and concern the dynamics of Wikipedia entries from a text linguistic point of view.
- Are you open to further your work in the framework of DARIAH lexical resources working group? YES
- VIAF: NO

3.

- Name: Krzysztof Nowak
- ESR: YES
- Fields of Interest (5 keywords): Medieval Latin, electronic lexicography, dictionary-encyclopaedia interface, non-lemmatic access
- Fields of Interest (up to 100 words): Krzysztof Nowak is at lexicographer at the Dictionary of Polish Medieval Latin (Institute of Polish Language, Polish Academy of Sciences). He obtained his PhD from the Jagiellonian University (*Comprehension of literary text in the ancient Latin commentaries to poetry*). He has been coordinating work on electronic dictionary and corpus of Polish Medieval Latin. In both projects he has been responsible for design, XML encoding, XSLT scripting and XQuery implementation of the dictionary interface (http://scriptores.pl). With Bruno Bon, he has been carrying out the project of the *WikiLexicographica*. Currently, he is working on Latin metaphors' description and Shiny-based application for simple corpus and dictionary statistics.
- Are you open to further your work in the framework of DARIAH lexical resources working group? I need more info in order to decide.
- VIAF: http://viaf.org/viaf/249518880

Image

1) It would be nice, if you could make sure we do have a picture representing you at our website (www.elexicography.eu).

2) Please add pictures and links representing your task group if you like! Please check on image licensing.

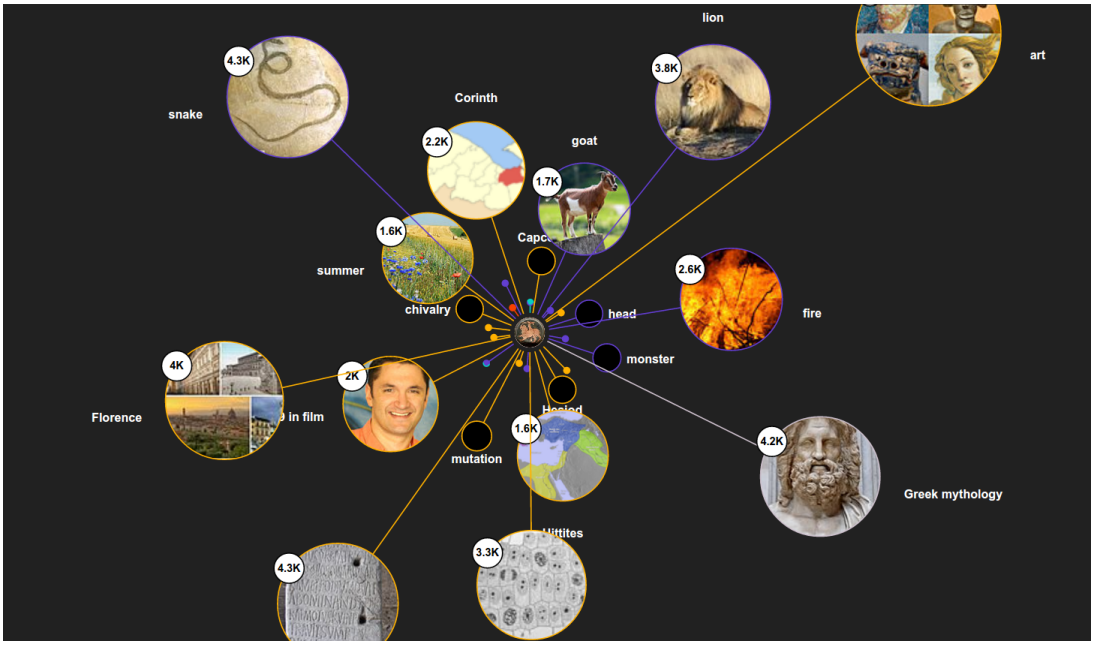

Image: Visualisation of the BabelNet entry *Chimaera* (available at: http://babelnet.org/synset?word=bn:00018403n&details=1&orig=chimaera&lang=EN): chimaera symbolises here an awkward combination of incompatible parts; the BabelNet represents dictionary-encyclopaedia interface and visualisation of lexicographic data.

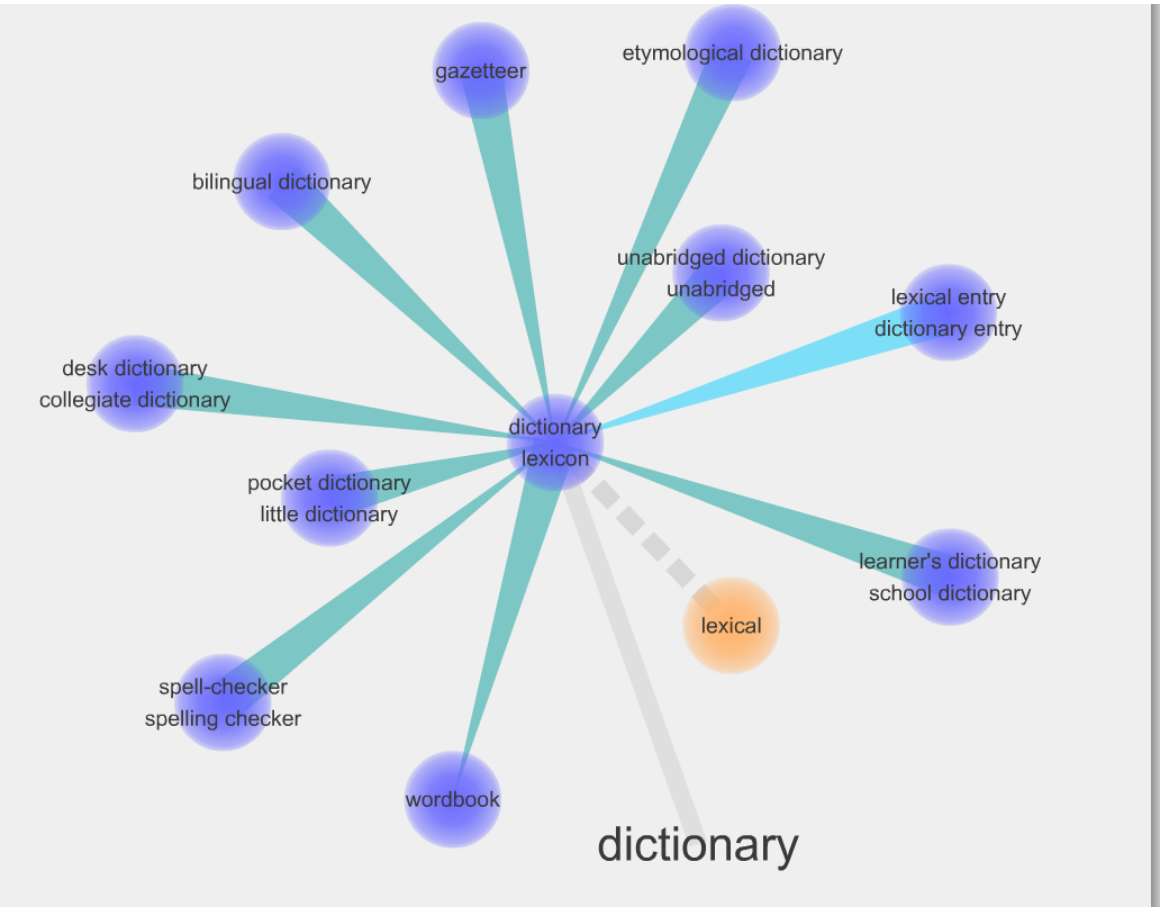

Image: graph representation of the WordNet entry for 'dictionary' (generated with *visuwords,* http://www.visuwords.com/?word=dictionary) represents both thesaurus work and lexicographic data visualisation

Krzysztof Nowak
krzysztofn@ijp-pan.krakow.pl