

**WG 4**

**“Lexicography and Lexicology from Pan-European Perspective”**

**Meeting Minutes**

**Date**

31.March/1. April 2016, 9:00-13:00/18:00

**Venue**

IULA- Institute for Applied Linguistics, Universitat Pompeu Fabra, Barcelona

**Sessions**

Session 1 – Lexical heterogeneity: March 31st: 9:00 – 13:00

Session 2 – Technical heterogeneity (Joint session with WG4): March 31st: 14:00 – 18:00

Session 3 – Social heterogeneity: April 1st: 9:00 – 13:00

**Storify:** <https://storify.com/msiemund/cost-enel-meeting-in-barcelona-wg4>

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Session 1 – Lexical heterogeneity: March 31st: 9:00 – 13:00**

The meeting was presided by the WG4 Chair, Eveline WANDL-VOGT.

**I. Opening**

Eveline opened the meeting and welcomed the attendees. The session chairs, Chris Mulhall and Geoffrey Williams opened the session, followed by their presentation.

**II. Presentations**

**1.** “LEXICOGRAPHY AND THE LANDSCAPE: PRACTICAL PHRASEOLOGY FOR PRACTICAL OUCOMES IN THE ENeL CONTEXT” (Chris Mulhall, Geoffrey Williams)

In their talk, Chris and Geoffrey explained their work on lexical variation and landscapes. They generally work on French (Geoffrey) and Italian (Chris), and also have an interest in Breton and Gaelic, but are not active in these fields at this point in time. Fog, understood as a metaphor, is the common connection theme. The types of variation addressed can be lexico-semantic, schematic, (meta-)linguistic or contextually-bound. They further addressed questions concerning the semantic relationship between lexical replacements, how to categorise them, and if variation is more common in VP or NP. Collocational networks & collocational resonance was mentioned as a means of displaying diachronic and synchronic variation. TEI is used to link existing dictionary sources. Interaction with other WGs is envisaged. Further, Chris is working on a development of learning tools for students to study lexical variation in French.

**2.** “HETEROGENEITY OF THE LEXICAL AND SEMANTIC ASPECTS OF THE DISTRIBUTION OF LATINISMS – CONTRASTIVE VIEW OF ENGLISH AND SLOVAK” (Adela Böhmerova)

Adela addressed heterogeneity in Latinisms in English and Slovak as to their presence, lexical form & semantic functions, systemic saturation and distribution, contrasting with the cross-linguistic situation. Interlinking heterogeneity is not dealt with. She mentioned that there was a recent boom in lexicography in Slovakia. Her project deals with extracted latnisims and possible comparison with other languages.

**3.** “HETEROGENEITY IN PLANT NAMES: LINGUISTIC HETEROGENEITY AND VOCABULARY” (Przemyslaw Debowiak & Jadwiga Waniakowa)

Przemyslaw and Jadwiga both talked about heterogeneity of plant names, addressing diachronic comparative analysis, aiming to reveal history of plant names, semantic motivation and origin. With the example of *eggplant*, they illustrate different origins of the name. They conclude that diachronic comparative analysis is important to plant names, and highlighted existance of numerous calques among plant names, and that dialectal names have various origins. Diachronic comparative analysis further allows to trace the ways of spread of various plant names in Europe.

**4.** “LEXICAL VARIATION IN MODERN DICTIONARIES: A COMPARISON BETWEEN GERMAN AND ENGLISH” (Anna Havina)

In her presentation Anna presented a case study on how modern dictionaries deal with lexical variation. A comparison between Englisch (New OED 1998; OED; Merriam-Webster online) and German dictionaries (Duden 1996, Duden Online, Österreichisches Wörterbuch 2012, Variantenwörterbuch des Deutschen 2004) was presented, where she looked at regional markings of the first 1000 entries in each dictionary. The existing frameworks for lexical variation are problematic: there is no standardized framework and often misrepresentations of variational forms. There is a need for a standardised framework - necessary information includes grammatical information, meaning, geographical distribution, register; arrangement of information.

**5.** “CROSS-LINGUISTIC PERSPECTIVES ON THE EUROPEAN LEXICON” (Wiebke Blanck, Esperanca Cadeira, Simeon Tsolakidis & Alina Villalva)

Alina and Wiebke presented their data from the *European roots dictionary*, with suggestions of a Pan European concept. They presented terms for emotions (anger) and colours (red) with examples for Portuguese and German. The Greek data is presented by Simeon in the next talk. They showed the structure of the entires and different roots for “red” in European languages. A future task includes discussing the actual dictionary structure.

**6.** “LINGUSTIC HETEROGENEITY AND VOCABULARY” (Simeon Tsolakidis)

Simeon talked about the Greek data pertaining to the root dictionary, with the main subset on ancient and hellenistic Greek. He explained the distribution of roots and associated words.

**7.** “CONNECTING EUROPEAN CULTURAL DIVERSITY: A CASE STUDY ON COLOUR TERMS – CHALLENGES IN A DIGITAL LEXICOGRAPHIC CONTEXT (Amelie Dorn, Eveline Wandl-Vogt & Thierry Declerck)

A digital humanities approach to lexicographical diversity was presented by Amelie. Colour terms here serve as an example case study. She addressed questions as to how diversity should be adressed in the lexicon, how languages could be structurally compared. With the help of DH tools it would be possible in Ontolex to model concepts and connect various sources. DH tools can offer possibilites to model and connect different sources and concepts. Challenges for digital lexicography remain, but could possibly be solved with DH tools.

**8. Wrap up & discussion:** Task Group “LandLex”: Aims, visions and contribution to the COST ENeL action

1. regarding organisation of task group LandLex and the ENeL Action in general:

A joint task group LandLex (based on the task groups lexical variation and european roots for colors and emotions) was established.

Geoffrey addressed the question as to how to proceed with the task group from now. Chris pointed out that a clear overview is needed on who is studying which language, as there are also overlaps between task groups. It was then discussed if papers or a proposal for an output would be more sensible, also in relation to the deliverables. Chris suggested that projects would attract more funding than papers, Alina also favoured the idea of projects, as there is still lots of work to be done and projects would give the chance to also involve PhD students. The idea of projects was generally favoured. A common Googledoc was put together by Alina and sent out to all interested participants of the meeting. Thierry mentioned that from a data point of view it’s important to have the data in structured way - Excel is the minimum to have it distributable. Wiebke raised the issue, that there is a need for a tight schedule until September. Chris added that also the deliverables and what can be done until the next meeting need to be fixed.

**Session 2 – Technical heterogeneity (Joint session with WG4):**

**March 31st: 14:00 – 18:00**

**I. Opening**

Eveline and Toma opened the meeting and welcomed the attendees. This was followed by a series of presentations.

**II. Presentations**

**1.** “OVERVIEW OF MODELS AND STANDARDS FOR SEMASIOLOGICAL AND ONOMASIOLOGICAL DATA” (Laurent Romary)

Laurent pointed out that there is nothing like one single model to represent lexicographical data. He showed different standards like TEI (P5), ISO (TC 37), e.g. LMF and TBX and W3C (SKOS, Ontolex). Standards are necessary for exchanging data, pooling heterogenous lexical data, interoperability between software components and comparability of results. One has to distinguish between onomasiolagical data (concept to term: TEI) and semasiological data (term to concept: TBX, LMF). He requires a stabilisation of good practices in TEI encoding, more convergence between TEI and ISO as well as a data delivery in flat models for the semantic web.

**2.** “MULTILINGUAL STANDARDS: XLIFF LESSONS” (Alexander O’Connor)

Alex started with the main interest: get the data out and get it back in again, even if it’s different. Therefore he presented different formats. First he introduced XLIFF, an XML-based format created to standardize the way localizable data are passed between tools during a localization process and a common format for CAT toolfiles. Its modules are: translation candidates (matches, mtc:), format style (fs:), metadata (mda:), resource data (res:), change tracking (ctr:), size restriction (slr:) and validation (val:). Further, he introduced TMX (Translation Memory eXchange), SRX (Segmentation Representation), TermBase eXchange and lemon core.

**3.** “PROPOSAL FOR LEXICAL INFORMATION MAPPING ARCHITECTURE (LIMA) FOR DICTIONARY EDITING AND PORTAL PUBLISHING” (Jens Erlandsen)

Jens introduced LIMA which combines lexicons from different sources with different formats. It indexes content (words, strings, etc.), finds parts of content based on match of a search criteria order, which are both, fixed and dynamically, and presents the content in one or more feasible ways. Thereby it takes care of by/in the idata format. LIMA changes the way of using XML by loosen/untie the complex burden of to the elements and their names and map parts of the information into attributes. Therefore it uses the attributes @type for representing the lexical information type, e.g. example, and @pres for defining the original layout.

**4.** “STANDARDS AND DIALECTAL DATA: THE EXAMPLE OF THE DATABASE FOR BAVARIAN DIALECT IN AUSTRIA (DBÖ)” (Jack Bowers, Melanie Seltmann)

Jack presented some technical aspects of the DBÖ. He gave an overview of characteristics and fields of TUSTEP and MySQL data as well of the conversion standards (XML, TEI, TBX\*, SKOS\*, ISO…) and a reusable infrastructure using LOD resources (babelfy service, DBpedia, wikidata, etc; GeoNames, ISOcat, GOLD, etc.). He also talked about his work on conversion and enhancement. It is planned to address specific needs of the data using a TBX-TEI hybrid model (cf. Romary, 2014). Further plans are to add something similar to the TBX <descrip>, <descGrp> elements for TEI dictionaries and to allow the head of an entry to be structured around a head concept.

**5. Wrap up & discussion:** Choosing the appropriate format(s) in a dictionary project workflow

Thierry pointed out that it is better to use SKOS, than SKOS-XL. Toma mentioned that the trouble with TEI is that it’s too flexible. An aim should be to define guidelines and stick to them. Laurent argued that parrallel examples are important for many lexicon objects, in TMX or XLIFF. Toma noted that the JSON community is very big, but it’s optional to use both, XML and JSON, but we don’t want lexicons to be written in JSON but in XML. He further pointed out to have LIMA as an approach for particular purposes (creating portal and putting different dictionaries together).

In general, the decision was taken to proceed with close collaboration with ongoing initiatives and projects, such as the DARIAH WG lexical resources (Laurent, Toma, Eveline) and H2020 project [PARTHENOS](http://www.parthenos-project.eu/) to make use of research infrastructures and work towards sustainability.

**6.** HANDS ON SESSION: From Encoding to Networking Dictionaries: Goals, Challenges and Best Practices

In the Hands on Session, several people offered overview and insight into ongoing lexicography projects from a interoperability point of view (with strong emphasis on TEI):

• Geoffrey Williams (FR): Standard TEI encoding of French historical dictionary

• Kseniya Egorova (IL): Dictionary of Russia

• Athanasios Karasimos (GR): Great Dictionary of the Greek Language

 (Το Μέγα Λεξικόν της Ελληνικής Γλώσσας)

• Martina Kramaric (HR): Miklošič, Lexicon Palaeoslovenico–Graeco-Latinum

• Ana Tešić (RS): The Dictionary of the Prizren Dialect

• Vladimir Benko (SK): Slovak Historical Dictionary in TEI: a feasibility study

• Jozica Skofic (SI): Slovene Dictionaries Network

• Kira Kovalenko (RU): Interlinking Russian Dialect Dictionary with Global Resource

**Session 3 – Social heterogeneity: April 1st: 9:00 – 13:00**

1. **OPENING**

Eveline opened the session. She pointed out that there is a gap between people making dictionaries and those who are no professional lexicographers. All participants brainstormed on possible connections between lexicography and Citizen Science. This was followed by a series of three talks.

**1.** “AN HOLISTIC VIEW OF CITIZEN SCIENCE” (Fermín Serrano Sanz)

Fermin presented a comprehensive overview of different aspects and examples of Citizen Science, as well as as different possibilites on how to work in CS. Inter alia, Socientize was mentioned as a project which supports different CS projects.

**2.** “CITIZEN SCIENCE IN THE CONTEXT OF RECENT DIGITAL HUMANITIES PROJECTS: AN OVERVIEW AND OUTLOOK” (Amelie DORN, Melanie SELTMANN)

Amelie presented different Digital Humanities projects which included some involvement of Citizen Science, most of them selected from different DH congresses in 2015. The presentation showed possible expectations and possibilites for the exploreAT! project (ACDH@ÖAW).

**3.** “SURVEY ON OPEN SCIENCE AND CITIZEN SCIENCE FOR ELEXICOGRAPHY” (Amelie DORN, Melanie SELTMANN, Sylvana KROOP, Maria SCHRAMMEL, Eveline Wandl-Vogt, Fermin Serrano Sanz)

Melanie presented the concept, process and content of the CS survey. With a discussion on possible CS integration in eLexicography and known CS and eLexicography projects, her presentation serves as the connecting point to the following WORLD CAFÉ Tables.

**4.** WORLD CAFÉ: Citizen science in lexicography

The talks were followed by a World Café session where the following 3 related topics were discussed:

● ENGAGEMENT – table chair: Fermín Serrano Sanz (ES)

● PARTICIPATION – table chairs: Amelie Dorn, Melanie Seltmann (AT)

● VISUALISATION – table chair: Roberto Theron (ES)

Results see (part of minutes): <https://docs.google.com/document/d/1P8gI8wU_oiGO-fj1fgAZ2meVG2_ACXfeMPQGExjo9z8/edit?usp=sharing>

**5.** Wrap up & discussion:

There was common understanding to proceed with the topic “citizen lexicography supported by visualisation” and improve towards concrete results during the next COST ENeL meetings.
The next steps to proceed with this is a joint proposal of several WG4-members for the DARIAH Public Humanities Call 2016 (submitted: 3 workshops - Lead: O´Connor: DCU [IE]). Further steps also include tight collaboration with the “citizen science infrastructure” [SOCIENTIZE](http://www.socientize.eu/). A piloting project on cultural concepts (bread) is established, presented at the European Citizen Science Congress (ECSA) 2016, which is open to be adapted for COST ENeL participants.

Citizen lexicography was considered as a relevant topic for further developments in scientific lexicography and science communication.

Amelie Dorn, MC Substitute AT
 Eveline Wandl-Vogt, Chair WG4