Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies

Amelie Dorn[1], Eveline Wandl-Vogt[1], Yalemisew Abgaz[2], Alejandro Benito Santos[3], Roberto Therón[3]

[1]Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Vienna, Austria [2]Adapt Centre, Dublin City University (DCU), Dublin, Ireland [3]VisUsal Group, Universidad de Salamanca, Salamanca, Spain
{amelie.dorn, eveline.wandl-vogt}@oeaw.ac.at
yalemisew.abgaz@adaptcentre.ie, {abenito, theron}@usal.es

Abstract

The world's indigenous languages and related cultural knowledge are under considerable threat of diminishing given the increasing expansion of the use of standard languages, particularly through the wide-ranging pervasion of digital media and machine readable editions of electronic resources. There is thus a pressing need to preserve and breathe life into traditional data resources containing both valuable linguistic and cultural knowledge. In this paper we demonstrate on the example of an Austrian non-standard language resource (DBÖ/dbo@ema), how the combined application of semantic modelling of cultural concepts and visual exploration tools are key in unlocking the indigenous knowledge system, traditional world views and valuable cultural content contained within this rich resource. The original data collection questionnaires serve as a pilot case study and initial access point to the entire collection. Set within a Digital Humanities context, the collaborative methodological approach described here acts as a demonstrator for opening up traditional/non-standard language resources for cultural content exploration through computing, ultimately giving access to, re-circulating and preserving otherwise lost immaterial cultural heritage.

Keywords: indigenous languages, cultural conceptualisation, data visualisation, semantic data modelling

1. Introduction & Background

In today's digital age, dialects, much like indigenous languages, are under considerable threat of diminishing as standard languages pervade the public domain as a means of communication, particularly in Western societies. The global decrease in indigenous languages and also in dialects or regional varieties of languages poses a considerable risk to maintaining not only linguistic knowledge diversity, but also cultural diversity and ultimately mankind's heritage. With this in mind, efforts by UNESCO (2002) have been made to sustain and foster dialogue around cultural diversity, including linguistic diversity. Similarly, educational minority- and under-resourced language initiatives have received fresh impetus and support over the past years (cf. Jones & Ogilvie, 2013). Although globalisation and the increased use of digital media as a means of communication have brought about a surge in standardisation across different fields of life, advances in computational capacities may at the same time also be exploited for maintaining knowledge diversity. Methods such as semantic modelling, the enrichment with (Linguistic) Linked Open Data (LOD)[1] or a combination of different computational processing and linking methods may enable sustainability, longevity and ultimately re-use of otherwise forgotten resources, contributing to the documentation of existing and new formations of diverse and rich knowledge networks.

In this paper, we thus showcase the potentials semantic enrichment (Section 3) paired with visualisation tools (Section 4) offer in revealing and giving access to unique traditional cultural knowledge and cultural conceptualisation contained within a non-standard language resource on the example of the Bavarian dialects in Austria, Europe, containing data from 1200 up to now, focusing on the early 1900s.

In what follows, we present work in progress and a first glimpse into the cultural conceptualisation contained in our digital resource (DBÖ/dbo@ema) (Wandl-Vogt, 2010), by looking at the original data collection questionnaires, which constituted the starting point of the collection at the time and which we therefore also take as our initial methodological case study. Our approach to tackling such resource with the methodology presented below is unique, and with our endeavour, we hope to serve as a demonstrator for other language resources of similar composition of which there are many around the world.

2. Non-standard Language Resource: the exploreAT!-case study DBÖ/dbo@ema

The questionnaire data described in this paper is part of a larger data collection, the Database of Bavarian Dialects in Austria [*Datenbank der bairischen Mundarten in Österreich*] (DBÖ) and related dbo@ema (Wandl-Vogt, 2010). The databases contain digitised data from questionnaires, related answers, as well as digitised entries from vernacular dictionaries and folklore literature. Apart from standard and non-standard German, the entries and dictionary excerpts in the database also dip into other languages such as Hungarian, Slovak, Slovene or Serbian, to name a few.

The questionnaire data we deal with here constitutes only a fraction of all data contained in the databases. It pertains to a dictionary project (WBÖ, 1970-), which aimed at capturing the German language spoken by local people from the early 20th century onwards in the area of the former Austro-Hungarian empire. Originally in analogue form (see Figure 1), the information from questionnaires and related answers (3.6 million individual digital entries) has undergone several stages of digitisation and is now available in XML/TEI formats (Schopper, Bowers & Wandl-Vogt, 2015). Apart from being a rich linguistic

[1] http://lod-cloud.net/

resource, the data also contain a wealth of historic cultural information of everyday life, e.g. customs, religious festivities, food, traditional medicine, professions, songs, etc. (cf. Wandl-Vogt, 2008). In addition, detailed information on persons (n=11,157) involved across several stages of data collection or processing is also available (cf. Piringer, Wandl-Vogt, Abgaz & Lejtovicz, 2017) as well as detailed geographical and location information (cf. Scholz, Hrastnig & Wandl-Vogt, 2018).

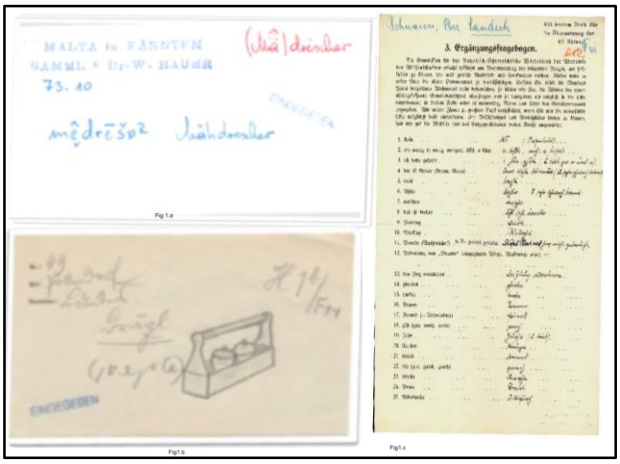

Figure 1: Left panels: examples of answer paper slips containing words, drawings, pronunciation, location and collector information. Right panel: an example of original analogue questionnaire and questions.

In the current Digital Humanities project *exploreAT! - exploring austria's culture through the language glass* (cf. Wandl-Vogt, Kieslinger, O'Connor & Therón, 2015), the cohort of these digitised data is now computationally processed from the aspects of cultural lexicography, semantic modelling, visualisation analysis and citizen science to access the traditional cultural knowledge and shed light on the knowledge system of the former society.

The questionnaires and their questions constitute the former starting point of data collection and are thus a key aspect in shaping the cultural content of the entire collection. Where projects with databases of similar content tend to entirely focus on the linguistic analysis of collected answers, we consider the questionnaires as an essential conceptual access layer to the collection. For this reason, we first aim to unlock the cultural concepts contained in the questionnaire questions, to extend the exploration to the remainder of the data in a second step.
The 120 questionnaires dealt with here thus concern three types: (1) Systematic Questionnaires [*Systematische Fragebogen*] (n=109), (2) Additional Questionnaires [*Ergänzungsfragebogen*] (n=9) and (3) Dialectographic questionnaires of the Munich and Vienna Dictionary Commissions [*Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen*] (n=2). Across these 120 questionnaires, we count a total of 24,382 questions asking for linguistic or cultural information or a combination of the two. The three questionnaire sets differ from one another according to form, content and purpose. In what follows we describe the application of semantic technologies (Section 3) as a first step in unlocking the cultural concepts.

3. Semantic Modelling of Cultural Knowledge Systems

For accessing the cultural content information in the questionnaires, the application of semantic modelling methods is essential. In general, various models have been designed to supplement semantics to original language resources (e.g. Chiarcos, Cimiano, Declerck & McCrae, 2013). These models don't only provide the meaning required to understand and correctly interpret these resources, but they also provide tools and techniques to effectively exploit their semantic information. To capture the semantics of the questionnaire questions, we followed a bottom-up approach (Noy & McGuinness, 2001). First, the original data collection methods, then the different types of questionnaires and questions were identified. This allowed gaining insights into the original approach taken and interpretation of the data. The questionnaires allow to aggregate and separate the resources based on the similarity of topics they address. The current semantic model (see Figure 2) captures the three types of questionnaires (systematic, additional and dialectographic questionnaires) based on the nature of the questionnaires and the type of information sought.

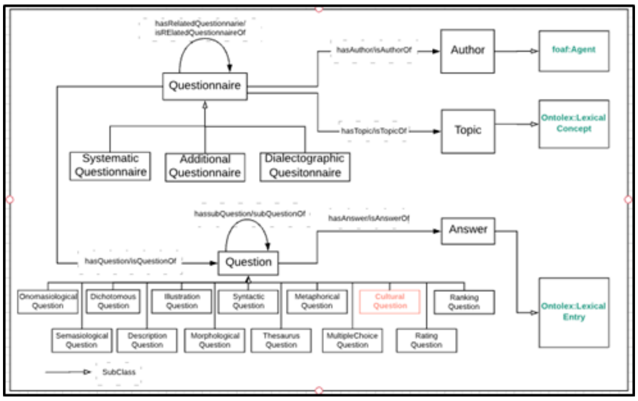

Figure 2: An initial semantic model of the questionnaires.

In addition, the model also captures authors and topics of individual questionnaires, where topics function as a means of aggregating the answers. Across the questionnaires, a total of 14 different types of linguistic questions were identified by relevant words or abbreviations contained in the questions themselves, including naming, definition, morphology, phonology, syntax, synonyms, etc. Cultural questions, on the other hand, contain significant information on representing and preserving the cultural identity of the communities and their language and were identified according to topics such as food, traditional medicine, games, songs etc., see Figure 3. In addition to the structure, we further capture patterns and examples of cultural questions which will later serve for the characterisation of questions of similar domains elsewhere.

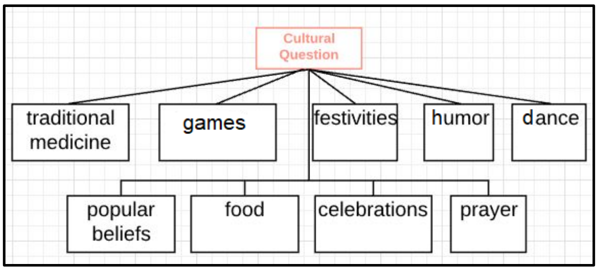

Figure 3: A data model of cultural questions.

Our model captures the original intent of the collection and provides a stable foundation for future interlinking with a variety of other resources, e.g. Europeana[2], Wikidata[3], Babelnet[4], Germanet[5] or relevant sources in the context of the co-creation approaches with user communities such as, for example, Topothek[6], Community Cooking[7] or Gastrosophie[8]. It further enables exploration of the data based on similar topics and structures and provides unique perspectives to navigate the entire collection by exploiting guided navigation to the answers of the questions. In addition, our semantic model provides the structure for the next stage of our data visualisation tools based on their semantic similarities.

4. Visual Discovery of Indigenous Cultural Knowledge Design via Concept Lights v.1.0

Data visualisation has become a key component of the Digital Humanities in recent years (cf. Benito, Losada, Therón, Dorn, Seltmann & Wandl-Vogt, 2016). By combining human-computer interaction techniques, psychology and graphic design, it can bring great insights to humanistic questions of the kind we discuss in this paper. In this line we developed the tool *Konzeptlichter/ Concept lights*[9] (Figure 4) that supports the visual exploration of the questionnaires introduced in Section 2.

The tool plays with lights and shadows to help the expert user to form a mental image of the structure of a single questionnaire by displaying common term associations and their disposition in the corpus. The foundation of the proposed linked-view system is an adjacency matrix representation, showing how many coincident terms are shared between different questions. These coincidences were detected and extracted in a previous data import stage where graph data structures were generated. Also in this step, questions were cleaned and stripped off stop words, leaving only semantically meaningful terms in each question, ready to be visualised. For example, the original question “Gesicht: Gesichtsrose, Rotlauf und andere Erkrankungen” is condensed as “Rotlauf, Gesichtsrose, Erkrankungen” in our approach. If any other question refers to the same terms, we consider these as semantically coinciding and those other concepts accompanying the matching terms are conceptually close and therefore relevant for the type of visual exploration we propose.

Two questions matching in two or more terms are considered a significant group. In turn, we use these groups to enable the exploration of the questionnaires. In Figure 4, groups are represented as coloured circles at the bottom left of the visualisation. Hovering over these groups, the terms are projected onto the matrix by illuminating the questions containing the terms.

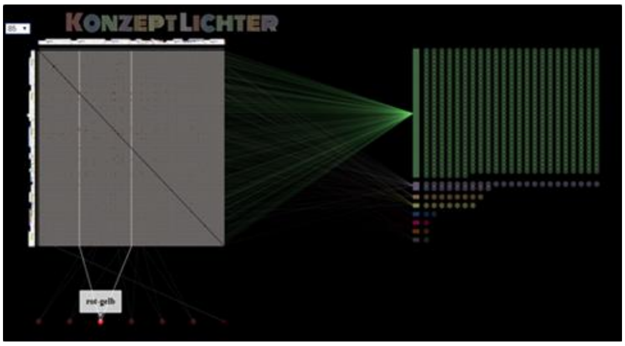

Figure 4: Konzeptlichter / Concept lights v1.0: Visual exploration prototype of words of single questionnaires. Selection of individual questionnaires (left) with the display of concept terms from questions (right).

Another way of exploring the questionnaire is enabled by the individual concepts found in the questions (see Figure 4). On the right part of the visualization, different terms found in the questionnaire are sorted by increasing order of importance. Terms found less often are placed at the top (green circles, appearing once, more abundant) whereas more common ones are moved to the bottom. (other colours, appearing two or more times, less abundant). An easy way of exploring the occurrence of terms within a questionnaire was enabled by applying the same highlighting effect described before when hovering over the individual circles representing the concepts. We also employ a magnifying effect that allows highlighting of all terms inside a group at once. These two annexe views are connected to the matrix by two-way channels, i.e. the highlighting effects also occur when selecting specific cells (questions) in the adjacency matrix. Whereas this tool is intended for the expert lexicographer with previous working experience with the questionnaires, we are designing new interaction paths tailored for the novel user that we expect to present in future research.

[2] https://www.europeana.eu/portal/ [accessed: 06.03.2018]

[3] https://www.wikidata.org/ [accessed: 06.03.2018]

[4] babelnet.org/ [accessed 06.03.2018

[5] www.sfs.uni-tuebingen.de/GermaNet/ [accessed: 06.03.2018]

[6] www.topothek.at/de [accessed: 06.03.2018]

[7] https://www.caritas-wien.at/stadtteilarbeit/ aktuelleprojekte/ community-cooking/[accessed: 06.03.2018]

[8] www.gastrosophie.at/ [accessed 06.03.2018]

[9] concept-lights.herokuapp.com [accessed: 06.03.2018]

5. Discussion & Future Work

Next steps in developing our semantic model contain digging deeper into the food domain as a case study for cultural conceptualisation: together with exploreAT!-community-groups (experts in various areas and lay-persons) a domain specific data-model for a thesaurus is co-designed, developed and evaluated.

Concept Lights v.2.0 aims to offer summarized insight into all questionnaires (not just one by one), incorporating the ontology model outlined in Section 3, which will in turn allow the proper contextualization of the displayed concepts and the retrieval of relevant content from the Semantic Web.

Concluding, exploreAT! aims to experiment with analogue and similar collection questionnaires to improve data modelling, visualisation as well as contextualisation of cultural and linguistic diversity as well as biodiversity and contribute to foster awareness about its wealth and its accessibility and documentation.

9. Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! project, carried out in a collaboration with the VisUSAL Group, Universidad de Salamanca, Spain and the ADAPT Centre for Digital Content Technology at Dublin City University which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

6. Bibliographical References

Benito, A., Losada, A., Therón, R., Dorn, A., Seltmann, M. & Wandl-Vogt, E. (2016) “A spatio-temporal visual analysis tool for historical dictionaries.” Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM). ACM. pp. 985-990. doi:10.1145/3012430.3012636

Chiarcos, C., Cimiano, P., Declerck, T. & McCrae, J. P. (2013). Linguistic Linked Open Data (LLOD). Introduction and Overview. In C. Chiarcos, P. Cimiano, T. Declerck & J. P. McCrae (Eds.), 2nd Workshop on Linked Data in Linguistics. Representing and Linking Lexicons, Terminologies and Other Language Data. Pisa, Italy, 23rd September 2013. Retrieved January, 17, 2018: http://www.aclweb.org/anthology/W13-5501.pdf

Jones, M.C. & Ogilvie, S. (Eds.) (2013) Keeping Languages Alive: Documentation, Pedagogy and Revitalization.CUP.

Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Technical, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Piringer, B.,Wandl-Vogt,E., Abgaz, Y., & Lejtovicz, K. (2017) Exploring and exploiting biographical and prosopographical information as common access layer for heterogeneous data facilitating inclusive, gender-symmetric research. In Wandl-Vogt, E. & Lejtovicz, K. Biographical Data in a Digital World 2017. A conference in the framework of the project APIS, 6–7 November 2017. Abstracts. [Wien]. doi:10.5281/zenodo.1041978

Scholz, J., Hrastnig, E. & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In P. Fogliaroni, A. Ballatore & E. Clementini (Eds.), Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017), (Lecture Notes in Geoinformation and Cartography (LNGC)). Basel: Springer International Publishing, pp. 275–282.

Schopper, D., Bowers, J. & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a database of dialects into a rich LOD resource. Retrieved January 17, 2018 from Text Encoding Initiative. Conference and members’ meeting 2015. October 28-31, Lyon, France. Papers: http://tei2015.huma-num.fr/en/papers/#146

UNESCO (2002) Universal Declaration on Cultural Diversity: a vision, a conceptual platform, a pool of ideas for implementation, a new paradigm. Cultural Diversity series, Vol.1 http://unesdoc.unesco.org/images /0012/001271/127162e.pdf [last access: 19.01.2018]

Wandl-Vogt, E., Kieslinger, B., O’Connor, A. & Theron, R. (2015). exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. In DHd2015. Von Daten zu Erkenntnissen. 23. bis 27. Februar 2015, Graz. Book of Abstracts.

Wandl-Vogt, E. (2008): Wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I Dinamlex) (mit 10 Abbildungen). In: Ernst, Peter (Eds.): Bausteine zur Wissenschaftsgeschichte von Dialektologie/ germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20.-23. September 2006. Wien. pp. 93-112.

[WBÖ] Wörterbuch der bairischen Mundarten in Österreich (1970–). [Dictionary of Bavarian Dialects in Austria] Bayerisches Wörterbuch: I. Österreich. Ed. by Österreichische Akademie der Wissenschaften. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

7. Language Resource References

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [*Database of Bavarian Dialects in Austria*] (DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010; Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [*Database of the Bavarian Dialects in Austria electronically mapped*] (dbo@ema). Wien. [Processing status: 2018.01.] https://wboe.oeaw.ac.at/dboe/indices/