Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 943–947
Marseille, 11–16 May 2020

Identification of Indigenous Knowledge Concepts through Semantic Networks, Spelling Tools and Word Embeddings

Renato Rocha Souza[1], Amelie Dorn[1], Barbara Piringer[1], Eveline Wandl-Vogt[1]
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences[1]
Sonnenfelsgasse 19, Vienna, Austria[1]
{renato.souza, amelie.dorn, barbara.piringer, eveline.wandl-vogt}@oeaw.ac.at

Abstract

In order to access indigenous, regional knowledge contained in language corpora, semantic tools and network methods are most typically employed. In this paper we present an approach for the identification of dialectal variations of words, or words that do not pertain to High German, on the example of non-standard language legacy collection questionnaires of the Bavarian Dialects in Austria (DBÖ). Based on selected cultural categories relevant to the wider project context, common words from each of these cultural categories and their lemmas using GermaLemma were identified. Through word embedding models the semantic vicinity of each word was explored, followed by the use of German Wordnet (Germanet) and the Hunspell tool. Whilst none of these tools have a comprehensive coverage of standard German words, they serve as an indication of dialects in specific semantic hierarchies. Methods and tools applied in this study may serve as an example for other similar projects dealing with non-standard or endangered language collections, aiming to access, analyze and ultimately preserve native regional language heritage.

Keywords: Digital Humanities; regional languages; lesser-resourced languages; knowledge discovery

1. Introduction

In this paper we present a technique we have developed for identifying dialectal words in a non-standard language legacy dataset (DBÖ [Datenbank der bairischen Mundarten in Österreich/Database of Bavarian Dialects in Austria]). This collection was analyzed in the context of the DH project exploreAT! (Wandl-Vogt et al, 2015). The project exploreAT! – exploring Austria's culture through the language glass evolved as a cross-disciplinary project at ACDH-OeAW since 2015. It brings together expertise from different disciplines and collaboration partners in the fields of cultural lexicography and open innovation (ACDH-OeAW, Austria), semantic technologies (ADAPT Centre, DCU, Ireland) and human-machine interaction via visualization (VisUSAL, Universidad de Salamanca, Spain) (Abgaz et al., 2018; Benito et al., 2016; 2018; Dorn et al, 2018). The project has given rise to, and enabled, the establishment of the Open Innovation Research Infrastructure (OI-RI) and exploration space at the Austrian Centre for Digital Humanities (ACDH-OeAW) at the Austrian Academy of Sciences

The project more generally aims at making the implicit, but to-date still partially unlocked cultural knowledge contained within a non-standard language legacy dataset accessible, connectable and reusable for different disciplines and actor groups. Even though heterogeneous data offers a wide spectrum of different kinds of uncertainties to be addressed, we here more specifically focus on the analysis of uncertainties in linguistic, spatial and temporal aspects of our legacy language collection.

The DBÖ collection is a valuable linguistic resource, containing regional, vernacular words that were spoken by the native community in an area of the former Austro-Hungarian empire. The collected words reflect the spoken language of the native community that carried the people's knowledge, culture and identity. For these reasons, the word collection can be considered indigenous, as it was part of their identity and captured and preserved traits of their customs. In addition, we can also consider it endangered, as linguistic variation, the preservation of local dialects and vernacular expressions are at risk due to globalisation. The prevalence of standard languages pose a threat to local language communities and the passing on of the wisdom and knowledge from previous generations.

2. The non-standard language collection

The exploreAT! project evolves around a digitised non-standard language resource of the Bavarian Dialects in Austria (DBÖ) and related dbo@ema [Database of Bavarian dialects electronically mapped] (Wandl-Vogt, 2008). A prominent part of the resource constitute digitised data collection questionnaires and related answers from paper slips, which initially also pertained to a dictionary project (WBÖ, 1963–) [Wörterbuch der bairischen Mundarten in Österreich/Dictionary of Bavarian Dialects in Austria], intending to capture the German language spoken by the local population, including also the compilation of a linguistic atlas of the local dialect geography (Arbeitsplan, 1912). This highly heterogeneous collection was amassed by the continuous application of a set of 24.403 questions, organized in 765 questionnaires to respondents in the area of the former Austro-Hungarian empire from the beginnings of the last century up to now. It captures the language and through it the culture of the local society. Besides capturing the local non-standard speech of the population, it contains a wealth of cultural information on detailed aspects of the former day-to-day life of the local population, including professions, customs, religious festivities, folklore medicine, food, etc.

Besides this dataset, the DBÖ collection further contains digitised information from excerpts of folklore literature, vernacular dictionaries or plant name and mushroom catalogues. The data then follows a lexicographic structuring consisting of lemmas, definitions, sources, time stamps, location information and a variety of other fields. In addition to this rich texture of linguistic, cultural and societal content captured, there is also detailed information available on persons (authors, collectors, editors) (Piringer et al., 2017), and spatio-temporal information (places, regions, GIS locations, etc) of the collection (Scholz, Hrastnig & Wandl-Vogt, 2018).

Tables 1, 2 and 3 present a numerical overview of a relevant sub-set (linguistic, location, time related) of the major entities contained in this non-standard language data collection, in two of the main current digital sources (XML/TEI files and MySQL database). Table 1 shows the time span of the entries in each of the main sources.

	XML/TEI files		MySQL DB	
time span of entries	**oldest**	**newest**	**oldest**	**newest**
year	1010	2008	1196	2012

Table 1: Numerical overview of temporal information on the entries. Source: the authors.

Table 2 illustrates the proportion of entries with and without spatial information.

	XML/TEI files		MySQL DB	
number of entries	2.416.499		65.839	
	with location	**without location**	**with location**	**without location**
	1.712.705 (71%)	703.794 (29%)	7333 (11%)	58.506 (89%)

Table 2: Numerical overview of entries with and without spatial information. Source: the authors.

Table 3 shows, for each one of the main databases, the number of entries with spatial information with a breakdown by level of location from the locations hierarchy.

	XML/TEI files		MySQL DB
Location level	**number of distinct locations per level**	**number of entries with locations**	**number of entries with locations**
• Bundesland	9	1.316.889 (55%)	-
• Großregion	32	1.296.722 (54%)	-
• Kleinregion	323	1.286.463 (53%)	415 (0,6%)
• Gemeinde	1.146	1.198.447 (50%)	3.058 (4,6%)
• Ort	1.145	1.198.447 (50%)	19.946 (30%)
• Ort (without associated Gemeinde)	24.788	395.186 (16%)	-

Table 3: Numerical overview spatial information per spatial level. Source: the authors

The DBÖ collection has undergone numerous levels of transformations since its beginning in the early 20th century (~1913) until today (2019). Originally, the collection was initiated with questionnaires and answers noted on individual paper slips. From these analogue forms, the cohort of the now digital data has evolved through several stages of digitisation and digital transformation through the years (see Figures 1 and 2) until reaching its current state; partly in XML/TEI formats (Schopper, Bowers & Wandl-Vogt, 2015) and partly as a MySQL database (dbo@ema) (cf. Wandl-Vogt, 2012).

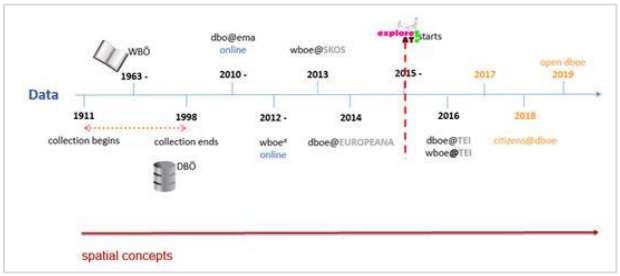

Figure 1: Timeline of the data transformation process relative to the beginning of the exploreAT! project. Image © amelie dorn, eveline wandl-vogt 2018.

During the turmoils of WWII the collection suffered some notable losses, thus the precise quantitative overview of the original material can only be speculated. In the first stage of digitisation (1993-2011), all available information noted on the paper slips (including, for example, headword, meaning, pronunciation, location, date, collector name, etc.) was manually entered by several data typists into TUSTEP (TUebinger System von TExtverarbeitungs-Programmen / Tuebingen System of Text Processing tools), resulting in ~2.43 million entries. Towards the end of this first digitisation process (2007), part of the TUSTEP data (auxiliary databases: persons database, literature database, plant name database; location data) were transferred to a MySQL database as part of the dbo@ema project (Wandl-Vogt, 2012). For the first time, different separate databases were joint; a geographic visualisation interface (maps) and GIS locations were added; and information and data were publicly made accessible and visible on the internet via a project website. From then, the heterogeneity of the data increased again with parts of the original data being still available in TUSTEP, another part having been converted to MySQL, and additionally newly digitised data being directly entered in MySQL. Following the dbo@ema project, the next step marked the transformation process towards the networked, open data realm. The conversion process of the remaining TUSTEP data to XML/TEI format was one of the first endeavours in exploreAT!, starting in 2015. Data entered in MySQL, however, remained unaltered at first. In the course of exploreAT!, the MySQL data and some of the XML/TEI data were converted to RDF and linked to LOD Cloud (2017-).

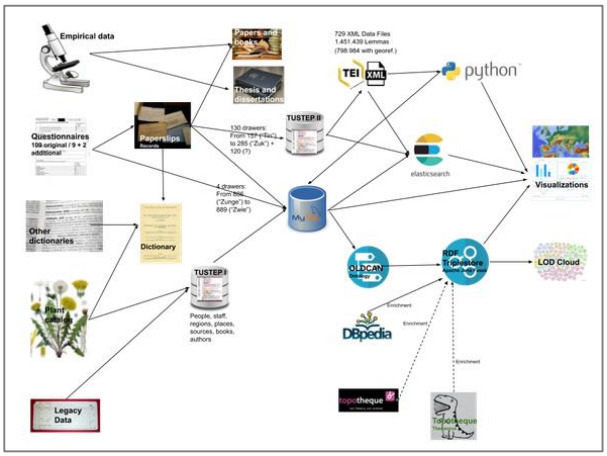

Figure 2: Overview of the data transformation process. Source: the authors.

In addition, the data is currently in the process of being enriched with lexical concepts and linked to DBpedia concepts. Figure 2 shows an overview of the data transformation process.

3. The data transformation process

Through intensive application of NLP and visualization techniques, the DBÖ collection has been studied under the aegis of the exploreAT! project. Some analysis include the lexical distribution of lemmas and linguistic borders (Figure 3), contextual words networks (Figure 4); classification and topic modelling, among others. Some of these results were presented in Rocha Souza et al., 2019.

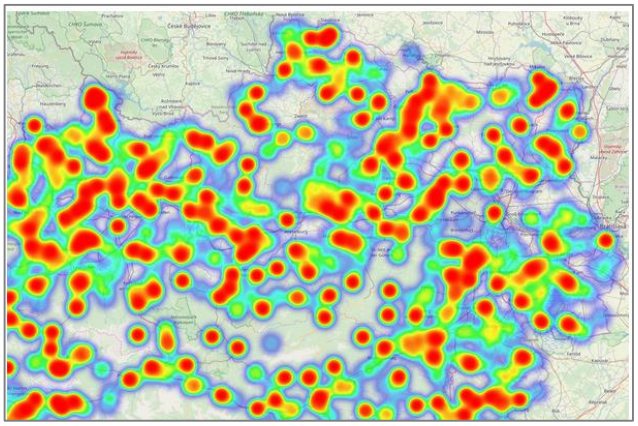

Figure 3: Part of the heatmap of the spatial concentration of collected concepts. Source: the authors.

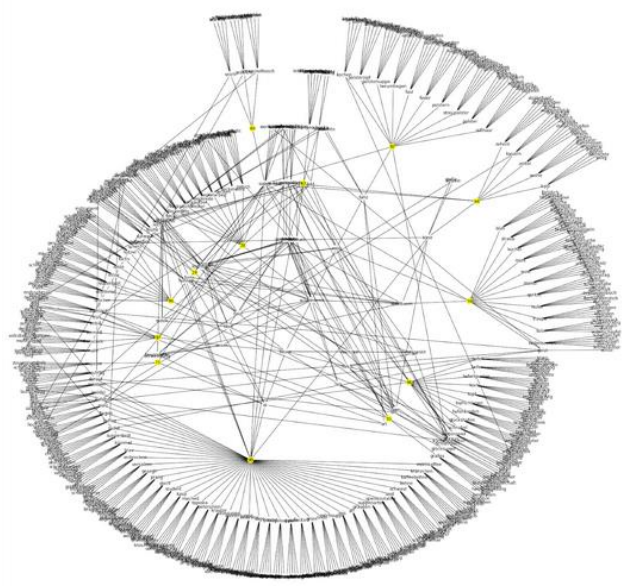

Figure 4: Networks of words appearing in the same contexts of questionnaires. Source: the authors.

Some effort in this project was also devoted to the identification of dialectal variations of words, or simply words that do not pertain to High German. This is important for the study of regional, vernacular and/or endangered languages. With the huge number of individual words in the collection (~31300, excluding the stopwords), it is difficult for manually identifying which ones would be candidates, for each one of the cultural categories analysed (events, sayings, beliefs, medicine, prayer, songs, humor, games, food, living organisms and dances).

First we have identified some common words from each of these cultural categories and took their lemmas using GermaLemma (Konrad, 2019). Then, using a word embeddings model (Mikolov et al, 2013), we have explored the semantic vicinity of each word. That led to some unexpected words associations, and also helped to harvest strongly correlated undictionarized words (see Table 4 and Figure 5).

word: Werkzeug	word: kopf	word: Gewerbe
('Werkzeuge', 0.78325855731964)	('sperrangelweit_offen', 0.623633146286601)	('Gastronomie', 0.69001352787017)
('Schraubenzieher', 0.73439121246337)	('angewurzelt', 0.59344947338104)	('Handel_Handwerk', 0.68979692459106)
('Schweissgeraet', 0.73391795158386)	('bombenfest', 0.58843773603439)	('Gewerbebetriebe', 0.68149685859680)
('Werkzeugen', 0.72254073619842)	('hinterm_Tresen', 0.5865101814270)	('Gewerbetreibende', 0.67372262477874)
('Akkuschrauber', 0.71929460763931)	('roten_Lettern', 0.58619940280914)	('Gewerbe_Industrie', 0.67345190048217)
('Werkzeugkoffer', 0.7122356891632)	('Haende_Hosentaschen', 0.58404660224914)	('Kleingewerbe', 0.66948682069778)
('Stemmeisen', 0.71130859851837)	('Praesentierteller', 0.57934200763702)	('Handel_Gewerbe', 0.66893702745437)
('Bolzenschneider', 0.701929509639)	('gluckst', 0.57766127586364)	('Einzelhandel', 0.66727524995803)
('Brecheisen', 0.69788682460784)	('splitterfasernackt', 0.57612943649291)	('Dienstleistung', 0.66460919380187)
('Bohrmaschine', 0.69774353504180)	('wackeligen_Beinen', 0.5749275684356)	('Dienstleistungsbetriebe', 0.64910757541656)

Table 4: Words highly correlated in the word3vec model. Source: the authors

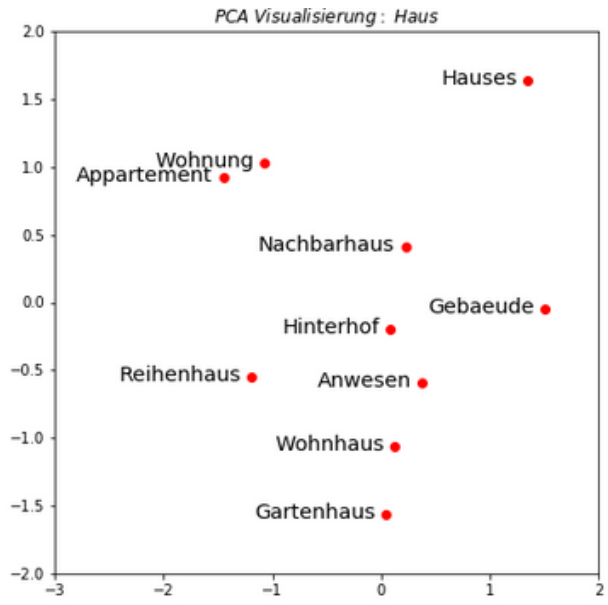

Figure 5: Principal components depicting the geometry of the vector space related to the word “haus”. Source: the authors.

The last automated step was to use the German Wordnet (Germanet) (Hamp & Feldweg, 1997) and the hunspell[1] to filter out dialect candidates. While it is known that none of

[1] "Releases - hunspell/hunspell". Retrieved 12 April 2017 – via GitHub.

these tools have a comprehensive coverage of standard german words, they serve as an indication of dialects in specific semantic hierarchies. As an example, applying the pipeline to the 899 words associated to the top Germanet categorie “artifakts” in the collection, we have found the 22 (2.44%) dialectal candidates:

- hantelet
- Barthlmä
- gaißkrack
- poganze
- maselicht
- gspusi
- linggelet
- hörndlbretze
- weinbeerschieben
- klungetzt
- nutschgeld
- potschamper
- geißhachsen
- kriecherln
- tristeler
- salasch
- perronkarte
- krigeln
- rinkerldürr
- ropfet
- gescherret
- kranewit

4. Conclusion

This paper presents merely the first exploratory approach of a technique to semi-automatically identify dialects in a compound traditional language legacy collection of the Bavarian dialects in Austria in specific cultural categories. Our study has demonstrated that with the application of existing semantic tools and methods, language concepts can be accessed and made connectable. At the same time, we have also highlighted the scarcity of non-standard, dialectal words in the existing semantic landscape, on the example of German resources. For indigenous languages, this has a number of implications: firstly, their linguistic and cultural wealth is still strikingly underrepresented in the available semantic networks. Secondly, accessing cultural and regional knowledge is key in preserving not only the language, but also traditional habits, processes and customs. These traditions otherwise run the danger of vanishing and being permanently eradicated by globalization, migration of populations predominantly veering towards the big cities and the looming loss of mankind’s tangible and intangible cultural heritage.

5. 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 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

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., & Way, A. (2018a). A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis. In McCrae, J.P., Chiarcos, C., Declerck, T., Gracia, J., & Klimek, B. (Eds.), Proceedings of the LREC 2018 Workshop "6th Workshop on Linked Data in Linguistics (LDL-2018)" (pp. 21–29). Miyazaki, Japan.

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., & Way, A. (2018b). Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers. Information, 9(12), 297:1–297:24. https://doi.org/10.3390/info9120297

Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912. Karton 1. Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch. Archive of the Austrian Academy of Sciences. Wien.

Benito, A., Losada, A.G., Therón, R., Dorn, A., Seltmann, M., & Wandl-Vogt, E. (2016). A Spatio-temporal Visual Analysis Tool for Historical Dictionaries. In García-Peñalvo, F.J. (Ed.), Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 985-990). New York, NY: ACM. https://doi.org./10.1145/3012430.3012636

Dorn, A., Wandl-Vogt, E., Abgaz, Y., Benito Santos, A., & Therón, R. (2018). Unlocking Cultural Conceptualisation in Indigenous Language Resources: Collaborative Computing Methodologies. In Soria, L., Besacier, L., & Pretorius, L. (Eds.), Proceedings of the LREC 2018 Workshop “CCURL2018 – Sustainable Knowledge Diversity in the Digital Age” (pp. 19–22). ISBN: 979-10-95546-22-1.

Hamp, B. & Feldweg, H. (1997). "Germanet-a lexical-semantic net for german." Automatic information extraction and building of lexical semantic resources for NLP applications. https://www.aclweb.org/anthology/W97-0802.pdf

Konrad, M. (2019). GermaLemma: A lem-matizer for German language text.https://github.com/WZBSocialScienceCenter/germalemma.

Mikolov, T., Sutskever, I., Chen,K., Corrado,G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119.

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. (Eds.), Biographical Data in a Digital World 2017. A conference in the framework of the project APIS, 6–7 November 2017. Abstracts. https://doi.org/10.5281/zenodo.1041978

Rocha Souza, R., Dorn, A., Piringer, B. and Wandl-Vogt, E. (2019). Towards A Taxonomy of Uncertainties: Analysing Sources of Spatio-Temporal Uncertainty on the Example of Non-Standard German Corpora.

Informatics 2019, 6(3), 34; https://doi.org/10.3390/informatics6030034

Scholz, J., Hrastnig, E., & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In Fogliaroni, P., Ballatore, A., Clementini, E. (Eds.), Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017) (pp. 275–282). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-63946-8_44

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

Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Bayerisches Wörterbuch: I. Österreich (1970–). Ed. by Österreichische Akademie der Wissenschaften. Wien, Austria: Verlag der Österreichischen Akademie der Wissenschaften.

Wandl-Vogt, E., Kop, C., Nickel, J., & Scholz, J. (2008). Database of Bavarian Dialects (DBÖ) electronically mapped (dbo@ema). A System for Archiving, Maintaining and Field Mapping of Heterogeneous Dialect Data for the Compilation of Dialect Lexicons. In Bernal, E., DeCesaris, J. (Eds.), Proceedings of the XIII Euralex International Congress (= Sèrie Activitats 20) (pp. 1467–1472). Barcelona: Documenta Universitaria.

Wandl-Vogt, E. (2012). Datenbank der bairischen Mundarten in Österreich @ electronically mapped. Projektbeschreibung. Retrieved from https://dboema.acdh.oeaw.ac.at/projekt/beschreibung/

[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. Retrieved from http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt

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/