Document Enrichment as a Tool for Automated Interview Coding

Ajda Pretnar Žagar,* Nikola Đukić,* Rajko Muršič‡

*Laboratory for Bioinformatics
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
ajda.pretnar@fri.uni-lj.si, nd1776@student.uni-lj.si

†Department of Ethnology and Cultural Anthropology
Faculty of Arts
University of Ljubljana
Zavetiška ulica 5, SI-1000 Ljubljana
rajko.mursic@ff.uni-lj.si

Abstract

While widely used in social sciences and the humanities, qualitative data coding remains a predominantly manual task. With the proliferation of semantic analysis techniques, such as keyword extraction and ontology enrichment, researchers could use existing taxonomies and systematics to automatically label text passages with semantic labels. We propose and test an analytical pipeline for automated interview coding in anthropology, using two existing taxonomies, Outline of Cultural Materials and ETSEO systematics. We show it is possible to quickly, efficiently and automatically annotate text passages with meaningful labels using current state-of-the-art semantic analysis techniques.

1. Introduction

Qualitative data coding is a well-established procedure in social sciences, particularly in sociology, cultural studies, oral history, and biographic studies. The technique is gaining ground in anthropology, where interview transcriptions abound. Ethnographic text coding can become a serious research technique, using existing ethnographic systematics, categories, vocabularies, and codes. Data coding facilitates the analysis of themes and close reading of the interview segments on each theme, which is one of the main analytical techniques of ethnographies in anthropology, be they computer-assisted or manual.

Computer-assisted qualitative data analysis (CAQDAS) is used to determine topics of interview segments, where the topics are not discrete but can overlap. The coder would normally define a codebook with the topics, then go over the text and label passages with corresponding tags. In the end, the coder can review selected topical passages, define topic co-occurrence, and extract a subset of documents on a specific topic.

Manual labelling can take a long time and requires a somewhat experienced coder to handle the tagging. However, we can construct an automatic pipeline for segment tagging due to the rapid development of natural language processing tools and language resources. The pipeline is built on the recent developments in ontology enrichment, which uses pre-defined ontologies (or taxonomies). Documents are preprocessed, and then the resulting tokens, typically words, are compared by similarity to tokens from the ontology. A simple approach is based on TF-IDF[1] transform. In contrast, the current state of the art uses graph models (i.e., YAKE) and word embeddings (Godec et al., 2021) for determining concept similarity.

Qualitative data coding is often based on grounded theory (Strauss and Corbin, 1997). The theory, which is more of an analytical approach, focuses on codes to emerge from the data (Holmes and Castañeda, 2014) rather than imposing them. Coding can also stem from a linguistic paradigm, especially semantic approaches, where text would be labelled based on the occurrence of words in it. The first approach still requires human input, while the second is based on unsupervised machine learning. Thus, having a general ethnographic taxonomy or classification scheme enables researchers to inductively elicit prevalent topics from the data rather than devising elaborate codebooks in advance. Our contribution is applying semantic annotation and ontology mapping to interview transcripts.

Semantic enrichment of documents means assigning conceptually relevant terms to documents or document segments. The procedure can include automatic keyword extraction, which identifies relevant keywords in the text (Bougouin et al., 2013; Campos et al., 2020) or relating existing lists of terms to texts (Massri et al., 2019). The latter can be either unsupervised or supervised. Unsupervised refers to the terms being scored by their similarity to the text and (multiple) terms assigned to each document if their similarity to the document is above a certain threshold. Supervised means the terms are used for document classification, where a document is assigned the most probable term.

In continuation, we propose a technique using unsuper-

[1]TF-IDF is a document vectorisation technique which uses word counts to describe documents, weighing them based on overall word frequency.

vised ontology enrichment to automatically label text segments with the corresponding topic labels. Automatic segment labelling uses existing (anthropological) taxonomies to label interview segments and thus assist researchers in navigating interview transcripts. The proposed technique doesn't apply only to anthropology – it could be used in any text analysis research. We use anthropology as a use case since the use of computer-assisted techniques is still somewhat rare in this discipline.

Finally, a short note on terminology. The term "ontology" is used in computer science to describe a structured hierarchical list of terms (Gruber, 1995), while in social sciences and the humanities, it means a branch of philosophy studying concepts of existence. In this paper, we use the term ontology in the former sense, sometimes referring to it as a taxonomy for clarity.

2. Interview transcripts

Interview transcripts are specific since they contain questions from the interviewer and answers from the interviewee. The transcripts are usually structured, with names or abbreviations denoting the speaker. If the interview is (semi-)structured, questions between different interviews will be very similar, if not identical. Moreover, interviewing a person often requires the interviewer to ask for clarification, affirm the interpretation of the answer or simply confirm (s)he understood what the interviewee said. Hence including questions in the analysis is often not a good approach.

Delineating between questions and answers depends on the structure of the digital document. A dedicated parser would consider new lines as segment delineations and names, pseudonyms, or initials as speaker identifiers. Ideally, the parser would consider the continuation of a reply, even when it was interrupted by the interviewer. But without a proper co-reference resolution for the given language (Žitnik and Bajec, 2018), it is difficult to determine such conceptual segments.

3. Related work

Back in 1983, Podolefsky and McCarty (1983) had an interesting idea - how about using computers to help us navigate numerous ethnographic notes and transcripts? Those were the days when most anthropologists stored their data on physical paper. Navigating such texts apparently required duplicating pages to store them under various categories. Nowadays, this is no longer necessary. Ethnographic data is often multimodal and predominantly stored digitally. It includes images, videos, and audio recordings along with the text. When navigating digital text data, one can easily use the "find" function to look for different text segments, while similar techniques exist for navigating other data types.

Nevertheless, organising interview data is not an easy task, and there are ways computers can help. Podolefsky and McCarty (1983) proposed developing coding categories for marking text passages. This is the precursor to modern qualitative data analysis software, such as NVivo, Atlas.ti, or MaxQDA. These, too, require a predetermined set of categories used for labelling the data.

Modern computer-assisted qualitative data analysis (CAQDAS) approaches don't require using punched cards with per-page summaries to navigate the text, as was the case in earlier times. They can quickly retrieve segments tagged with the specified tag. MacMillan and Phillip (2010) use a semi-anthropological approach to better gauge the connection between venison price and cull effort. They conduct in-depth interviews with stalkers, people employed by the British estates that hunt wild game, and analyse the interviews with NVivo. They use the qualitative data from the interviews to corroborate the quantitative findings – deer hunting is deeply rooted in tradition and seen as a sport rather than economic activity.

Researchers studying sensorially-charged biographic experiences in Turku, Brighton, and Ljubljana defined the main categories with a larger list of subcategories. Coding only the translated transcripts and using Atlas.ti, they extracted similarly charged testimonies related to different sensations, for example, sounds (Venäläinen et al., 2020) or smells (Bajič, 2020).

Most commonly, CAQDAS is used in discourse analysis. Hitchner et al. (2016) analyse discourses on bio-energy to elicit key metaphors used to create common imaginaries. Using this approach, they were able to identify three discursive units that guide the bio-energy narrative. Cuerrier et al. (2015) identified 134 categories referring to climate change in 46 interviews conducted with the Inuit population in Nunavik. Next, they created ordinal and binary matrices describing the change in quantity and the presence or absence of topics. They used various statistical approaches to determine whether different communities of Nunavik differ in terms of knowledge of climate change. Both papers retrieve popular taxonomies created by people under study.

Discourse analysis is also prominent in Schafer (2011), who uses Atlas.ti to analyse over 30 in-depth interviews with secular funeral officiants called "funeral celebrants" in New Zealand. The author identified key conceptual categories in funeral celebrant ethnographies, specifically the narratives on connection, identity, and personalisation of funeral practices.

CAQDAS can also be used to retrieve relevant text passages. Yilmaz et al. (2019) conducted 30 interviews with highly educated Turkish-Belgian women to determine the factors affecting their marriage choices. They stem from grounded theory and use predetermined codes for the first round of coding, then refine and enhance their codebook later. With iterative codebook improvements, they determined women's decisions and the driving factors behind them, for example, the structural and general constraints in marriage choices.

Conversely, Wehi et al. (2018) do not use CAQDAS software but instead observe raw word frequencies in Māori oral tradition. They collected ancestral sayings called whakataukī and identified references to animal extinctions in the data.

It is interesting to note that many contributions using quali-quantitative text analysis were published in the Human Ecology journal, which testifies to the (still) marginal use of these methods in anthropology. Ideally, we will see many more journals willing to publish such research and

more researchers ready to use these tools in practice.

Longan (2015) expresses the sentiment to perfection: "There is room for innovation in the creation of technological aids to facilitate mesoscale qualitative online research that lies between massive data sets and small qualitative studies. Though the major qualitative software suites have improved over time, much of the process is still tedious and requires hours of snorkelling and coding by hand." First, he explicitly points to the nor-big-nor-small issue of many contemporary anthropological studies. Even organising just thirty interview transcripts can be complicated, let alone a hundred records. Yet one hundred records can hardly be described as "big data" requiring "big tools". There's a need for a mid-level tool to help organise the data in a time-efficient way. Second, he points to the issue of coding by hand, which takes time and effort from the researcher. Third, he identifies an opportunity for technological innovation for qualitative data analysis that surpasses modern qualitative analysis software.

Previously, ontology enrichment for labelling text passages was used predominantly in biology and medicine (Bifis et al., 2021). In social sciences and the humanities, automated segment labelling was expressed as more of a wish rather than a reality (Hoxtell, 2019). In contrast to CAQDAS, ontology enrichment provides a way to automatically label large amounts of text in a short period of time. At the same time it enables relating interview transcripts to existing domain-specific ontologies. Our contribution showcases automated interview segment labelling with existing ontologies, thus providing a practical example of how machine learning can support ethnographic analysis.

We propose an approach using ontology enrichment from computer science to help organise and structure interview transcripts, fieldwork notes, and archive data. The three-fold example described below is a prototype for machine-assisted data coding, which uses standard anthropological taxonomies, such as the Outline of Cultural Materials (Bernard, 1994, p. 519-528), or more local and specific ethnographic taxonomies, related to the European ethnology studies of the so-called folk or traditional culture (Kremenšek et al., 1976), to label text passages.

4. Ontologies as codebooks

Instead of pre-defining codebooks for manual coding, we propose to use existing anthropological taxonomies to automatically label the data. One such well-established taxonomy, which we call "ontology" in text mining, is the Outline of Cultural Materials. Human Relations Area Files is a non-profit research organisation whose aim is to foster cross-cultural research (Melvin, 1997). One of its key achievements is the establishment of several databases that contain previous cross-cultural research. The database entries, such as ethnographic reports, are indexed using the Outline of Cultural Materials (OCM), an ethnographic subject classification system developed by Murdock and colleagues (Murdock et al., 1969; Ford, 1971).

The taxonomy is designed in a decimal classification system, similar to the librarian Universal Decimal Classification. Its main categories start with Orientation (10), Bibliography (11) and Methodology (12), and end with Socialisation (86), Education (87) and Adolescence, Adulthood, and Old Age (88). The categories are still very general, so more specific categories must be coded additionally.

Ethnographic systematics (ETSEO) is derived from continental ethnographic practices, mostly interested in traditional culture of the European peasantry. Its taxonomy is hierarchically extensive, starting with the essentially defined material, spiritual and social culture categories. Since the taxonomy was designed for museum archives, the most detailed is the field of material culture, subdivided on as many levels as necessary, and taxonomy in general fits folk taxonomy and practices. Spiritual culture is further divided into general categories comprising folklore, ritual practices, and art-related activities. Less detailed is the so-called "social culture" field containing festivities in a calendar year, celebrations of live events, and communal activities, practices, and rules. This system is much more detailed but, at the same time, only partly decimally classified and only somewhat comparable to the OCM taxonomy. It was designed for classical archive work and is now only partially accepted as a digitised taxonomy.

OCM's main aim was to facilitate searching the large database of ethnographic entries and organise basic information on ethnic and social groups. Hence it is easy to extend the idea of an ethnographic classification system to a codebook – each entry represents a concept relevant to describing a culture. One could use the well-defined system with descriptions of categories to automatically tag text passages with relevant ethnographic concepts. For example, if the passage describes using outdoor toilets, the corresponding codes should be "744 Public Health and Sanitation", "515 Personal Hygiene", "336 Plumbing", and "312 Water Supply". Besides already existing taxonomies for ethnographic materials (OCM and ETSEO), it is useful to produce native or folk taxonomies as "a description of how people divide up domains of culture, and how pieces of a domain are connected" (Bernard, 1994, p. 386). Automated accurate tagging would enable quickly retrieving relevant parts of the text on the one hand and observing dominant topics and their inter-relatedness on the other.

5. Document enrichment

Analysis of interview transcripts would normally include labelling documents or interview segments with corresponding codes, identifying topics/codes, observing their frequencies in the corpus, and retrieving interview segments for a given topic/code. We show how to perform these tasks in a visual-programming data mining tool Orange (Demšar et al., 2013). Workflow (as seen in Figure 1) for replicating the analysis is available online (Pretnar Žagar, 2022b) along with a Slovenian translation of OCM ontology (Pretnar Žagar, 2022a). The corresponding data are not publicly available due to privacy issues.

5.1. Data and preprocessing

To demonstrate how contemporary ontology enrichment and semantic analysis approaches can be used in anthropology, we are using interview transcripts from twenty interviews on smart buildings (Pretnar and Podjed, 2019). The interviews are in colloquial Slovenian and describe

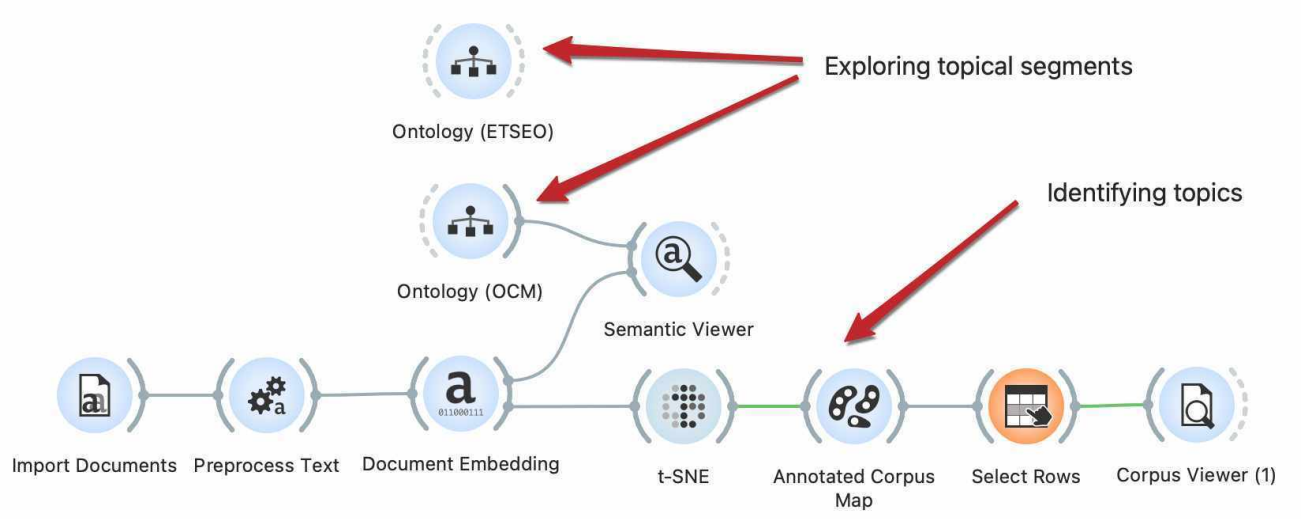

Figure 1: Workflow for ontology enrichment and extracting interview topics from annotated visualization.

the experiences and struggles of faculty staff with a smart building. The interview is segmented into questions and answers. Each answer represents the utterance and constitutes a single document in the final corpus resulting in 1126 data instances. The metadata includes the question, the interviewee, and the interview date.

Tokens are constructed by passing the text through the CLASSLA pipeline for non-standard Slovenian. Then, lemmas and POS tags are retrieved, and only nouns and verbs are kept for the analysis. Tokens are used to compute document embeddings, a mean-aggregation of word embeddings based on fastText models (Bojanowski et al., 2017). We tried simple lowercasing, Lemmagen lemmatization (Juršič et al., 2010) and stopword removal for preprocessing, but the results were not as informative (they mostly contained generic verbs, such as to have and to go, discourse particles and fillers). Moreover, while SBERT embeddings generally perform better due to their context-parsing abilities, they produced worse results in the t-SNE visualisation. Specifically, fastText identified a group of segments with short, unspecific replies (i.e., "Yes.", "Uh-huh."), while SBERT did not.

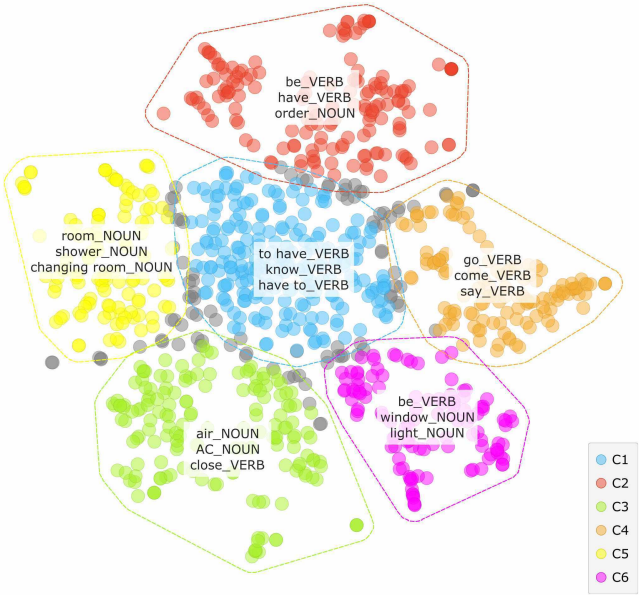

Figure 2: t-SNE document map with annotated semantic groups.

5.2. Identifying topics

Generally, the researchers will know which topics the corpus covers because often, they will be its creators. In the case of interviews, the researcher is likely also the interviewer who guided the interview based on research questions. However, ethnographic narratives often take unexpected turns or focus on unforeseen details, which the researcher can uncover by coding the data and iteratively refining the codebook. Alternatively, one can use document maps, where segments with semantically similar content will lie close together.

To semantically represent the content of interview segments, we will pass them to document embedding. The procedure will take the words (tokens) identified in preprocessing and find their vector representation. The representation models the meaning of the words in a way that relates "king" to "prince" and "queen" to "princess". Once the embedding of each word is retrieved, words from the document are aggregated into the mean document vector.

This numeric representation will be used to plot a t-SNE document map, where segments with similar content will lie close to each other[2]. But a bare map is not very informative on its own. Hence, we added Gaussian mixture models to identify groups of segments and retrieve their characteristic words (Figure 2). The procedure identified segments referring to air quality (green cluster), lighting (magenta cluster), room descriptions (yellow cluster), and so on.

[2]In t-SNE, we selected a larger group of segments for annotation. There was a smaller group of 121 segments representing short replies, such as "yes", "no", and "I don't know".

5.3. Exploring topical segments

Ontologies can be used to enrich interview segments by measuring how similar given ontology terms are to each segment. Automatic identification of segments helps researchers quickly identify relevant parts of the interview.

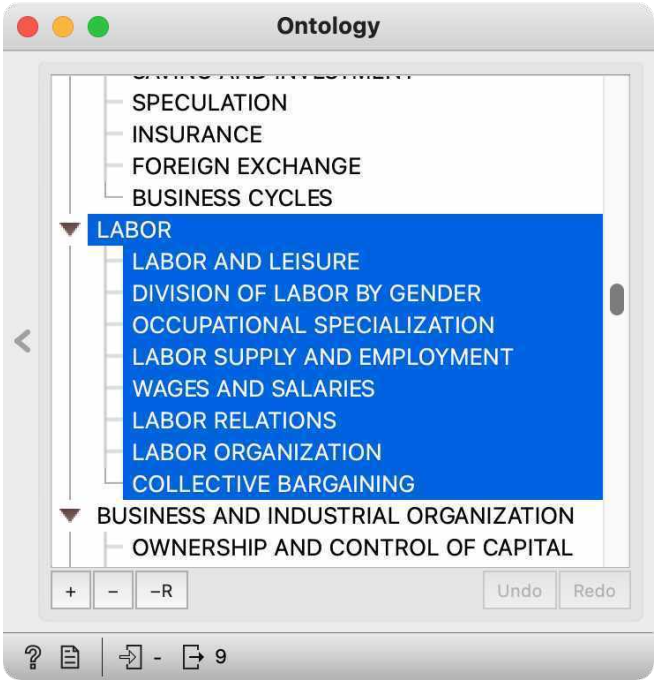

Figure 3: Selecting a part of the OCM ontology referring to work ("delo") and work-related terms.

For example, we can look for "delo" (orig. 350 equipment and maintenance of buildings) and its child terms from the OCM ontology in the corpus (Figure 3). Selected terms from the ontology are used for semantic annotation of interview segments.

Semantic annotation scores each segment by how similar its sentences are to the input terms, using SBERT embeddings (Reimers and Gurevych, 2019). SBERT was used because it specialises in sentence embeddings and considers word context. Ideally, this procedure identifies passages talking about work-related topics, including breaks, employment, paychecks, and work relations. One can sort the results by either the overall segment score, an aggregate of all sentence scores, or by matches, which counts how many input words appear in the segment.

Here, we show the latter option, namely displaying the segments with the most matches. We have selected all the segments matching any of the input terms and highlighted them (Figure 4). Ontology enrichment successfully identified segments discussing the office environment, research work, work routine, schedules, weekend work, etc.

5.4. Assigning terms to segments

The final goal of any automated coding system would be to return a corpus with assigned codes. We prototyped a procedure that uses the above technique of semantic scoring to identify the code with the highest score for each segment. We decided on a 0.6 cosine similarity threshold for the code

At work, not so much. More so in the afternoon.

It mostly works. But one sometimes has to fix things.

Mostly, yes.

Yeah, sure. I mean PhD students for sure.

From the office?

Both. If we are in the meeting at the same time, then absolutely, if we're on a break, the same.

For research, for teaching preparations, for meetings with the students, with PhD students, with anyone who comes to me. Purely for work purposes.

That works.

Figure 4: Annotating text segments with a part of ontology referring to work ("delo").

to be assigned, which resulted in segments that did not have a corresponding code. After loading the corpus, we remove all the interview segments without any codes. We retain 252 segments with codes and observe their frequencies. The results are somewhat promising but with some obvious errors (Figure 5).

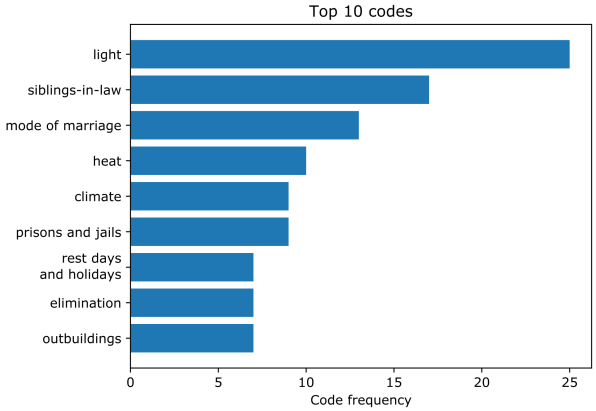

Figure 5: Top 10 codes identified in the corpus. While some are plain wrong, most are quite accurate and useful.

The most frequent code is "luč" (light), which is indeed a very prominent topic in the corpus. Then the results get a little strange. The two next topics are "svaki in svakinje" (brothers and sisters in law) and "tipi porok" (marriage types), which are not among the interview topics. The errors are likely caused by the multilingual SBERT model used for word embedding, which sometimes cannot distinguish between South Slavic languages. For example, it considers the Slovenian slang term "ratal" (succeeded) as "war" based on its similarity to the Serbian "rat" (war).

However, there are some quite relevant topics among the top ten codes, for example, "toplota" (warmth), "podnebje" (climate), "dnevi počitka in dela prosti dnevi" (rest

days and holidays), “stranska poslopja” (outbuildings), and “bivališča” (dwellings). Clicking on a label, for example, “toplota” (warmth), outputs text segments discussing the interviewees’ attitude to temperature regulation. With a few steps, the researcher can identify and extract interview segments discussing a specific topic and read them to better understand the context of these segments and which subtopics the respondents deem relevant. For example, the texts on temperature regulation mostly refer to difficulties with adjusting office temperature.

The system could be improved with specifically developed language resources for non-standard Slovenian. Nevertheless, even in its current imperfect form, it can be a useful tool for semi-automated coding, where the researcher can manually adjust the suspicious/incorrect codes.

5.5. Comparison to ETSEO taxonomy

While the OCM taxonomy is widely recognised in the anthropological community, the ETSEO taxonomy is strictly regional. The project Ethnological Topography of Slovenian Ethnic Territory (ETSEO) began in 1971 by a large group of Slovenian ethnologists led by Slavko Kremenšek. The project entailed the development of the questions based on ethnological systematics, ethnographies of Slovenian towns and cities (18 in total), and detailed ethnographies on a specific topic. The taxonomy is a result of the first part of the project, namely the questions and detailed ethnological systematics. The ETSEO questions were published between 1976 and 1977 in twelve books, including the introductory volume with reports of ethnological institutions (Kremenšek et al., 1976) and eleven volumes of topical presentations and suggested questionnaires. The series served as a theoretical and practical guide for ethnographic fieldwork (Ravnik, 1996).

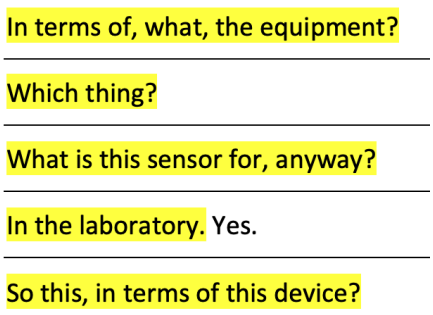

Figure 6: Matches for ETSEO entry “technical knowledge”.

ETSEO taxonomy contains 53 areas of ethnographic interest. Still, it lacks explicit hierarchy, although it follows the classical division of ethnographic material for the so-called folk culture: material (volumes I to V), social (volumes VI to VIII) and spiritual (volumes IX to XI). A rough hierarchy could be formed from the eleven books in which these questions were published, but the books lack hypernyms. Hence, we will use this as a flat taxonomy. There are fewer relevant areas to choose from than in the OCM. However, looking for “tehnično znanje” (technical knowledge) returns relevant interview passages (Figure 6).

The ETSEO taxonomy is less useful than the OCM taxonomy. This is due to the somewhat outdated nature of the questions, which were based on the main foci of Slovenian ethnology and were less relevant for anthropology. They are missing some key contemporary areas of anthropology, namely media, urban areas, internet communities, and migration. Nevertheless, the taxonomy could be extremely useful for older ethnographic texts and, with some updates, even for contemporary materials.

6. Conclusion

Anthropology can greatly benefit from the recent developments in text analysis. Ontology enrichment, along with other data exploration and visualisation methods, is a useful tool providing an overview of the collected data.

In the time when anthropologists are using larger corpora (Culhane and Elliott, 2016), when data is created online for many different purposes (Wang, 2012), and when anthropologists use online platforms to store raw ethnographic multimedia data (Przybylski, 2021), it is of utmost importance to store and later archive data meaningfully, using relevant classification and coding systems. It is even more important in archival work, which is no longer just an additional part of anthropological research, supplementing ethnographic fieldwork, but is becoming highly relevant for digital aspects of our lives.

Updating taxonomic systems is an urgent task for anthropologists. However, using existing taxonomies to explore and visualise data already benefits the analytic process, especially in re-studies and comparative research. Classical anthropological coding of ethnographic material is no longer possible, so automated coding is the first step to expanding the range of anthropological data analysis. However, in the absence of specialised word embedding models for Slovenian (SBERT is currently multilingual and conflates South Slavic languages), the approach does not yet achieve the accuracy of a human annotator.

While automated coding, particularly for languages with fewer language resources, still has a long way to come to be comparable to human input, it facilitates data exploration and extracting general topics from the text. Ontology enrichment tools support the iterative analytical process of ethnography. They provide a starting point for forming new research questions, enhancing existing ones and can be easily repeated on new data.

Many improvements could be made to improve automated coding for the Slovenian language:

- Developing a Slovenian-only sentence transformer used in semantic search.
- Re-writing transcripts in standard Slovenian or further improving CLASSLA to handle slang terms and non-standard Slovenian.
- Implementing co-reference resolution for Slovenian to resolve issues with indirect references in text, further clarifying the exact content of the document.

While these improvements would greatly enhance coding capabilities for Slovenian, they are, for the most part,

available for larger languages, thus already enabling similar research.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022–2027) and the DARIAH-SI research infrastructure.

8. References

Blaž Bajič. 2020. Nose-talgia, or, olfactory remembering of the past and the present in a city in change. *Ethnologia Balkanica*, 22:61–75.

H Russell Bernard. 1994. *Research Methods in Anthropology: Qualitative and Quantitative Approaches*. Sage Publications, Thousand Oaks, London, New Delhi.

Aristeidis Bifis, Maria Trigka, Sofia Dedegkika, Panagiota Goula, Constantinos Constantinopoulos, and Dimitrios Kosmopoulos. 2021. A hierarchical ontology for dialogue acts in psychiatric interviews. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, page 330–337, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Alain Cuerrier, Nicolas D Brunet, José Gérin-Lajoie, Ashleigh Downing, and Esther Lévesque. 2015. The study of inuit knowledge of climate change in nunavik, quebec: a mixed methods approach. *Human Ecology*, 43(3):379–394.

Dara Culhane and Denielle Elliott. 2016. *A Different Kind of Ethnography: Imaginative Practices and Creative Methodologies*. University of Toronto Press, North York, Ontario, Canada.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, et al. 2013. Orange: data mining toolbox in python. *The Journal of Machine Learning Research*, 14(1):2349–2353.

Clellan S Ford. 1971. The development of the outline of cultural materials. *Behavior Science Notes*, 6(3):173–185.

Primož Godec, Nikola Đukić, Ajda Pretnar, Vesna Tanko, Lan Žagar, and Blaž Zupan. 2021. Explainable point-based document visualizations. *arXiv preprint arXiv:2110.00462*.

Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5-6):907–928.

Sarah Hitchner, John Schelhas, and J Peter Brosius. 2016. Snake oil, silver buckshot, and people who hate us: metaphors and conventional discourses of wood-based bioenergy in the rural southeastern united states. Human Organization, 75(3):204–217.

Seth M Holmes and Heide Castañeda. 2014. Ethnographic research in migration and health. *Migration and Health: A Research Methods Handbook*, pages 265–277.

Annette Hoxtell. 2019. Automation of qualitative content analysis: A proposal. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 20.

Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Slavko Kremenšek, Vilko Novak, and Valens Vodušek. 1976. *Etnološka topografija slovenskega etničnega ozemlja. Uvod. Poročila.* Raziskovalna skupnost slovenskih etnologov, Ljubljana.

Michael W Longan. 2015. Cybergeography irl. *Cultural Geographies Special Issue - New Methods in Cultural Geography*, 22(2):217–229.

Douglas Craig MacMillan and Sharon Phillip. 2010. Can economic incentives resolve conservation conflict: the case of wild deer management and habitat conservation in the scottish highlands. *Human Ecology*, 38(4):485–493.

M Besher Massri, Sara Brezec, Erik Novak, and Klemen Kenda. 2019. Semantic enrichment and analysis of legal domain documents. *Artificial Intelligence*, page 2.

George Peter Murdock, Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W. M. Whiting. 1969. *Outline of Cultural Materials*. Human Relations Area Files, New Haven.

Aaron Podolefsky and Christopher McCarty. 1983. Topical sorting: A technique for computer assisted qualitative data analysis. *American Anthropologist*, 85(4):886–890.

Ajda Pretnar and Dan Podjed. 2019. Data mining workspace sensors: A new approach to anthropology. *Prispevki za novejšo zgodovino*, 59(1):179–196.

Ajda Pretnar Žagar. 2022a. *OCM ontology - Slovenian*. Figshare. https://doi.org/10.6084/m9.figshare.19844107.v1.

Ajda Pretnar Žagar. 2022b. *OCM ontology enrichment*. Figshare. https://doi.org/10.6084/m9.figshare.19787065.v1.

Liz Przybylski. 2021. *Hybrid Ethnography: Online, Offline, and in Between*. Sage Publications, Los Angeles; London; New Delhi; Singapore; Washington DC; Melbourne.

Mojca Ravnik. 1996. Način življenja slovencev v 20. stoletju. *Traditiones*, 25:403–406.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Cyril Schafer. 2011. Celebrant ceremonies: life-centered funerals in aotearoa/new zealand. *Journal of ritual studies*, 25(1):1–13.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded Theory in Practice*. Sage.

Juhana Venäläinen, Sonja Pöllänen, and Rajko Mursic. 2020. The street. The Bloomsbury Handbook of the Anthropology of Sound.

Tricia Wang. 2012. The tools we use: Gah- hhh, where is the killer qualitative analysis app? http://ethnographymatters.net/blog/2012/09/04/the-tools-we-use-gahhhh-where-is-the-killer-qualitative-analysis-app/.

Priscilla M Wehi, Murray P Cox, Tom Roa, and Hēmi Whaanga. 2018. Human perceptions of megafaunal extinction events revealed by linguistic analysis of indigenous oral traditions. *Human Ecology*, 46(4):461–470.

Sinem Yilmaz, Bart Van de Putte, and Peter AJ Stevens. 2019. The paradox of choice: Partner choices among highly educated turkish belgian women. *DiGeSt. Journal of Diversity and Gender Studies*, 6(1):5–24.

Slavko Žitnik and Marko Bajec. 2018. Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 6(1):37–67.