Building a Language Technology Infrastructure for Digital Humanities: Challenges, Opportunities and Progress

Dana Dannélls[1] and Daniel Brodén[2,3]

[1] Språkbanken Text/Department of Swedish
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se
[2] Dept. of History of Ideas, Literature and Religion
[3] Swe-Clarin/Centre for Digital Humanities
University of Gothenburg, Sweden
daniel.broden@lir.gu.se

Abstract. Språkbanken Text, a research unit at the University of Gothenburg, forms part of the National Language Bank of Sweden and is the main coordinating node of Swe-Clarin, the Swedish national CLARIN node. During the past years, Språkbanken Text has been actively engaged in a number of humanities and social sciences related research projects. This engagement has primarily concerned the development of new resources, methods and tools to accurately process large amounts of digitized material, in addition to interfaces for visualizing the materials, making them easily accessible for further analysis. The activities within Swe-Clarin have been essential for the progress and the success of this work. In this paper we present what was required from Språkbanken Text in order to meet the expectations of researchers from the humanities and social sciences. We discuss some of the challenges this work involves and describe the opportunities this field brings with it and how these opportunities could help to progress the work of Språkbanken Text toward building a language technology infrastructure that supports interdisciplinary research.

Keywords: Digital Humanities · Language Technology · National infrastructure

1 The humanities research problem

Researchers coming from digital humanities and its adjacent fields are in need of sophisticated tools, interfaces and materials to help them answer their current research questions and form new ones.

As a prominent institution in the field of language technology (LT), Språkbanken Text, a research unit at the University of Gothenburg, is on a regular basis approached by researchers from the field of humanities but also the social sciences with various methodological questions and feature requests. Consequently, over the years, Språkbanken Text has nurtured a growing interest in projects combining a language technology

approach with interpretative methods associated with other fields. While such collaborative efforts have contributed to positioning Språkbanken Text centrally in the Nordic digital humanities community, they have also provided new challenges.

A key question for Språkbanken Text today is how to continue developing language technology tools and services and at the same time answer to the needs and wishes of the wider research community. As critical commentators have stressed, while research infrastructure projects are vital for the introduction of computational methods for humanities research, there needs to be a strong focus on intensifying and allowing "creative forms of humanities research for the 21th century, not their replacement by science as a hegemonic form of knowledge creation" [2, page 84].

In this paper, we address the specific question of what is required from a language technology infrastructure, such as Språkbanken Text, in order to truly meet the expectations coming from the humanities and social sciences (HSS) researchers. More specifically, what actual steps and activities should be considered and above all prioritized in this context? We also address the question of how these measures contribute to helping language technology to progress with the work towards building a national infrastructure that supports truly interdisciplinary research.

We begin by presenting ongoing efforts at Språkbanken Text, where a major focus has been on developing and enhancing the lexical and corpora resource repositories and improving the tools to access them.[4] The enhancements and improvements are partly the results of close collaboration and dialogue with, among others, historians, librarians, rhetoricians, and literary scholars. Because the lexical and corpora resources of Språkbanken Text have been developed in interdisciplinary collaboration with researchers from different fields who possess different perspectives and knowledge about which pieces of information are important to encode, it has proved to be a challenging task to apply computational methods that are able to access the resources uniformly, as well as to develop tools that allow users without substantial programming skills to use the resources and to present the analysis results.

After this, we turn to the user involvement activities of Swe-Clarin, the Swedish CLARIN node, of which Språkbanken Text is the main coordinating partner, focusing in particular on interdisciplinary pilot projects and user workshops. In order to spread awareness of the research potentials of LT tools, Swe-Clarin and Språkbanken Text has initiated a series of one-off projects, in which we have collaborated with researchers from various fields of the humanities. Besides the research results and their dissemination, one vital outcome has been further insights into the needs and challenges posed by scholars from other disciplinary domains.

1.1 Språkbanken Text – a Growing Swedish Infrastructure

Språkbanken Text carries research on language technology for historical and modern Swedish, and develops a research infrastructure to support research in linguistics and other disciplines in the humanities, social sciences, and medicine.

Språkbanken Text was established by a governmental decree as a national centre of computational lexicography (then named Logoteket) in 1975. Through the centre,

[4] All the resources and tools at Språkbanken Text are CC-BY licensed.

computational linguists and corpora users, in and outside of Sweden, have been able to access linguistic and statistical data about a wide range of texts in Swedish for more than four decades [10]. During the 2000s, the work at Språkbanken Text rapidly expanded towards the development of high quality language technology tools. The first version of the widely used corpus tool Korp [6] was launched in 2011. Today, Korp is used in Sweden as well as in several other countries and has been modified to support other languages, among others by the Centre for Language Technology in Copenhagen and by the Language Bank of Finland, Kielipankki. Other tools, also named after birds and other animals, have followed, including the annotation tool Sparv [4] and the lexical platform Karp [5]. Furthermore, since 2014 Språkbanken Text is officially the main coordinating node of Swe-Clarin, the Swedish national node in the European Research Infrastructure Consortium CLARIN (Common Language Resources and Technology Infrastructure). As a part of, and a so-called B-centre, the CLARIN ERIC consortium, Språkbanken Text offers primarily HSS researchers access to a range of tools and corpora as well as services and knowledge on a sustainable basis.

2 The solution

The work at Språkbanken Text is a collaborative effort between researchers, so called experts, and system engineers. The daily work primarily comprises: (a) collection and development of digital resources; (b) development of tools and metadata standards to annotate and classify the material and link it to other resources; (c) implementation and maintenance of interfaces to visualize and navigate between the resources; and (d) analysis of the content of the resources by applying state-of-the-art methods. All of these are essential tasks for meeting the expectations coming from HSS researchers [18]. In the following, we draw on some them with emphasis on how the work at Språkbanken Text is progressing.

Resources As the result of mass digitization of historical and literary data, a wide range of new electronic resources have been added to Språkbanken Text's repository. One recently added resource is the Kubhist corpus of Swedish historical newspapers, spanning from 1645 until 1926, and containing around 5.5 billion tokens [1]. It is a rich resource for conducting diachronic studies and for examining various research questions concerning language and culture. Even though this new Kubhist has been digitized by the National Library of Sweden by applying state-of-the-art OCR technology, it contains many OCR-errors which pose new challenges to the annotation tools of Språkbanken Text. While there is an ongoing initiative to improve the OCR processes of Swedish newspapers [8], accessing the data in meaningful ways still requires specific methods and competencies to improve the performance of the analysis [9]. Meanwhile, regardless of the OCR-errors, the material could be explored to answer important research questions about language properties such as: What are the meanings and associations of different concepts at different points in time? What spelling variation is most prominent during a certain time period?

Two other complex and large resources that have been added to the resource repository of Språkbanken Text are the Twitter and Flashback corpora. Social media data is extremely diverse and the sensitive information it contains makes it compelling to

explore, for example, for sociologists or media scholars who are interested in finding value-laden, harsh words directed towards journalists and other public figures. Typical research questions this material could potentially provide answers to are: How have the statements made by a certain person affected his/her public persona? What kind of attention has a certain public figure attracted and through what social media activities?

A valuable lexical resource, the Swedish sentiment lexicon [16], was developed and added to the repository of lexical resources of Språkbanken Text. The resource is forming the first step towards the creation of a full-fledged sentiment analysis for Swedish. It has been proven in particular useful for analyzing opinions on social media.

Språkbanken Text has ongoing research collaborations with the Department of Historical Studies, the Department of Languages and Literature and the Centre for Digital Humanities at the University of Gothenburg. In relation to these collaborations, Språkbanken Text develops and maintains several databases, including the Biographical Dictionary of Swedish Women, a historical database containing biographies of women who have made significant contributions to Swedish society and culture,[5] and the NordiCon database, containing medieval Nordic personal names attested in Continental sources [20].[6] Språkbanken Text provides easy access and editing possibilities of these databases though Karp, Språkbanken Text's lexical infrastructure [5]. Drawing on the data from these databases one could ask research questions such as: How has the professional life of women in Swedish societies progressed the last couple of hundred years? What are the naming preferences among different religious communities? Are there any naming traditions that characterizes certain time periods?

Annotations Språkbanken Text's annotation pipeline is available through the tool Sparv [4] and over the years the pipeline has been increased with several annotation layers. Sentiment annotation is an example of an annotation layer that has been added to the pipeline [15], as one outcome of the Swe-Clarin activities. Furthermore, Sparv has been empowered with a user-friendly interface with possibilities to upload a digitized material in various formats and to automatically enhance it with different annotations such as sentiment, name entity, part-of-speech. The annotation results are displayed through the interface and can also be downloaded. Sparv's user-interface has been proven valuable for humanities scholars, and by following their requests it is constantly evolving.

Analysis Korp is Språkbanken Text corpus infrastructure tool. The tool comes with a graphical web user interface and functionalities for importing, annotating and exporting corpus resources. Users can explore the resources they are interested in by, for example, compiling statistics over the data and visualize the results with graphs, geographical maps or "word pictures".

Word picture is a functionality in Korp for visualizing word correlations in a text. Word picture has been widely used to explore various research questions related to conceptual transformations as well as to specific issues concerning equality or discrimination of old people or people with disabilities. Some examples are: When did certain words come into use, and when did they disappear? What words are used to describe men and women today compared to the early 20th century?

[5] https://skbl.se/en

[6] https://spraakbanken.gu.se/karp/tng/?mode=nordicon&lexicon=nordicon

Word picture has also been used to help scholars of political science to identify what different topics are prominently discussed by different political parties [14], answering research questions such as: What are the most frequent questions discussed under a certain election year? How does the public interest in certain topics change over time?

Visualization of geographical locations is another example of a functionality that has been added to generate maps marked with the locations of the place names that are are mentioned in the text. It can be exploited for answering research questions such as: Which geographical locations are mentioned in the work of a particular author and in what contexts? What do we know about the the etymology and history of a particular place name, and how has its geographical associations changed over time? What place names no longer exists? Such questions have received a great interest over the past years, especially from literary scholars. However, because of annotation errors, spelling variation, and lack of normalization, the automatic analysis unfortunately still fails to recognize many significant place names [3]. As a result, it is not always possible to explore the research questions to a satisfactory extent [13]. What adds to this insufficiency is the lack of information about the accuracy of the results, and a functionality that allows users to inspect the material through close reading.

Recently, Korp was enhanced with a functionality that enables researchers to conduct manual analysis of video transcriptions. This was developed as part of a collaboration with a group of researchers at the Institute of Language and Folklore in Sweden, who are working on a project entitled Interaction and Variation in Pluricentric Languages (IVIP). The new functionality allows, for example, anthropologists or sociologists to closely study how interpersonal relationships are expressed in institutional conversations in Sweden and Finland in the domains of service, education and healthcare. Possibilities for close readings of the text and for analyzing it directly on the screen is a functionality that has proved to be highly appreciated.

3 The collaboration experience

As mentioned above, Språkbanken Text is a also driving force in the user involvement activities of Swe-Clarin, that aim at spreading awareness of what research possibilities language technology tools can offer HSS scholars who use text and speech as primary research data. Digital humanities projects are often conducted with either strong data science or humanities bias [17]. Thus, one of the central means for spreading awareness of what research possibilities Swe-Clarin's digital tools can offer and at the same time meet the requirements of HSS researchers is collaborative pilot projects. Swe-Clarin has initiated a series of interdisciplinary one-off projects, in which scholars and technicians from Språkbanken Text collaborate with HSS researchers from various disciplinary fields. The projects have concerned, among other things, rhetoric, second-language acquisition and political science [13, 19].

A recent historical study [11] used text mining of the large newspaper corpus Kubhist to study the emergence of terrorism in the Swedish newspaper discourse from late 18th to early 20th century. The aim was partly to evaluate prior research claims regarding the historical connotations and ideological contexts of terrorism before World War I. The study could confirm what we already knew about the meanings of terrorism from

the historical record, but also go beyond common historical wisdom in pointing to the diversity of the meaning of terrorism in the period. For example, the word pictures of ”terrorism” and ”terrorist” had more prominent attributions of so-called state terrorism than expected. However, it should be noted that even with the large quantity of text in the Kubhist corpus, many of the queries returned only a few hundred hits, which does not allow for solid generalizations.

Besides the specific research results and their dissemination, one vital outcome of the user involvement activities of Swe-Clarin have been concrete insights into the needs of HSS researchers and into the further development of Språkbanken Text's resources to meet such demands. Apropos the Kubhist corpus, the researchers involved are looking forward to continuing the work on historical discourses on terrorism in Sweden by looking at an expanded version of Kubhist. The new data-set is five times as large and have better OCR quality [1], a factor which affects search accuracy [12]. The updated Kubhist corpus also allows for tests of the usefulness of diachronic word embeddings for studying the change in meanings of terms related to terrorism and other concepts over time. Furthermore, the pilot project on the conceptual history of terrorism has pointed to the need of supporting comparative analysis with versions of Korp installed in other countries. As the lead researcher has an interest in both Swedish and Finnish history, the project's focus has now been extended to include a comparative analysis of a Finnish newspaper corpus (in the Swedish language). This requires content search involving both the Swedish Korp tool and the version used by the Language Bank of Finland, similar to the more general federated content search that Språkbanken Text already provides to CLARIN.

Another important user involvement activity of the CLARIN endeavor is the organization of user workshops. Although Språkbanken Text and other members of the Swe-Clarin consortium provide various digital tools and materials for humanities scholars to use in their research, the resources are hardy used by all who would benefit from them [21]. The Swe-Clarin workshops (nicknamed “Swe-Clarin on tour”), which are held at universities and memory institutions around Sweden, give hands-on training in the use of some of the key tools of Språkbanken Text, such as Korp and Sparv, and in the formulation of productive research questions.

However, the aim of the workshops is not only to promote a wider utilization of Swe-Clarin's resources, but also to get user-feedback. Consequently they are carried out with an active interest in the research perspectives that the participants bring to the table. Similar to the collaborative pilot projects, the workshop dialogues have given us access to a body of perspectives and experiences that have provided insight into current limitations of our tools, from bugs to functionality gaps. A recurring feedback from the workshops has been the need for document-level access to the text materials provided by Språkbanken Text. While researchers in language technology have limited interest in, for example, cultural and historical content aspects of our corpora, such aspects are elemental to many researchers in the humanities. Thus, Språkbanken Text has initiated extensive work to meet the need for document-level access to texts, the most prominent example being the further development of the Korp tool [7]. Looked at from this perspective, as much as the workshops are about providing hands-on training in the use of our tools and in the formulation of productive research questions, they are

also about gathering feedback that helps us to develop the tools and data-sets for the needs of HSS scholars.

4 Conclusions and recommendations to educators

The various collaborative projects tied to the activities of Språkbanken Text and Swe-Clarin, in addition to the feedback from user workshops, has given us access to a body of perspectives and experiences that have provided valuable insights into current limitations of our tools and resources and contributed to their further development. Above all, they made us aware of the importance of integrating interdisciplinary perspectives and cooperations at different levels of digital resources and tools development.

As the results of extensive Swe-Clarin activities, Språkbanken Text has become aware of a range of desirable functionalities through input from political science scholars, historians, sociologists, rhetoricians, and ethnologists, who appreciate our infrastructure but have needs and research questions that can not currently be answered because of various technical and computational limitations. While some of the limitations were outlined in Section 2, we would like to round-off with highlighting some others that are currently lacking but will hopefully be added to the infrastructure of Språkbanken Text.

Real-time data analysis Social data contains huge amounts of up-to-date information and should therefore be analyzed the moment the data becomes available. Språkbanken Text does not yet have any support for uploading the data and for performing real time analysis. Moreover, currently, real-time calculations over a large amount of material take too long to be included in an interactive interface, making it difficult for investigating synchronic and diachronic questions which requires large data-sets.

Semantic similarity analysis We have so far been able to significantly improve word picture functionalities thanks to feedback and requests from researcher and users. Still, identifying words with semantic similarity and comparing their usages in positive and negative sense is one example of a user request that has been suggested to enhance the usability of word picture. Another suggested improvement of the word picture function is to generalize it to allow for comparison between the different annotations.

Upload ones own material and access it through the interfaces Researchers outside Språkbanken Text are still unable to upload their own material and explore the annotation results through our interfaces. Materials that have subsequently been analyzed should also be made available for download in a number of different formats for import in other tools than those of Språkbanken Text. This would allow researchers to further explore the materials, but also to continue working with their own research data, for example, by running analysis against existing collections at Språkbanken Text. In addition, while functionality for automatically identifying and annotating lexical entities in a text is available in the annotation pipeline of the tools of Språkbanken Text, it could be improved to allow for searches across internal and external resources. The merits of offering such a platform is two-sided; not only will HSS researchers benefit from it, but it will also be beneficial for Språkbanken Text by increasing our resource repository with valuable data-sets. Because there are various formats for how different materials are stored, we are interested in what formats different researchers prefer.

Normalization of articles Normalization is a necessary next step for drawing objective conclusions about the analysis results, and for answering questions such as how many authors are represented in the data or how many texts have been written by a specific author. However, this is a user request that requires substantial technical efforts, but becoming aware of the types of normalization that researchers are primarily interested in might help the developers at Språkbanken Text with finding plausible technical solutions.

As pointed out earlier in this paper, the daily work at Språkbanken Text is carried out by a group of experts and research engineers who are collaborating actively to take in user requests and generalize them to find suitable solutions and appropriate methods that could be made available through our interfaces. Thus, knowing the needs and wishes of users is an important step towards making the infrastructure accessible to a larger group of HSS researchers. Conducting pilot studies with researchers about their specific research interests is also a step toward spreading wider awareness in the HSS community about the potentials of language technology in research as well as further identifying user needs, and in extension, toward strengthening the capabilities of the infrastructure of Språkbanken Text and the National Language Bank of Sweden for all of their users.

Acknowledgements

The research presented here is supported by Språkbanken Text and Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant agreement 821-2013-2003).

References

1. Adesam, Y., Dannélls, D., Tahmasebi, N.: Exploring the Quality of the Digital Historical Newspaper Archive KubHist. In: Proceedings of the 4th Conference of The Association Digital Humanities in the Nordic Countries (DHN) (2019)
2. Berry, D., Fagerjord, A.: Digital Humanities: Knowledge and Critique in a Digital Age. Polity Press, London (2017)
3. Borin, L., Dannélls, D., Olsson, L.J.: Geographic visualization of place names in Swedish literary texts. Digital Scholarship in the Humanities **29**(3), 400–404 (2014)
4. Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., Schumacher, A.: Sparv: Språkbanken's corpus annotation pipeline infrastructure. In: SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016 (2016)
5. Borin, L., Forsberg, M., Olsson, L.J., Uppström, J.: The open lexical infrastructure of Spräkbanken. In: Proceedings of LREC 2012. pp. 3598–3602. ELRA, Istanbul (2012)
6. Borin, L., Forsberg, M., Roxendal, J.: Korp — the corpus infrastructure of Spräkbanken. In: Proceedings of LREC 2012. pp. 474–478. ELRA, Istanbul (2012)
7. Borin, L., Tahmasebi, N., Volodina, E., Ekman, S., Jordan, C., Viklund, J., Megyesi, B., Näsman, J., Palmér, A., Wirén, M., Björkenstam, K., Grigonyte, G., Gustafson Capková, S., Kosiński, T.: Swe-clarin: Language resources and technology for digital humanities. In: Digital Humanities 2016. Extended Papers of the International Symposium on Digital Humanities Växjö, Sweden. Edited by Koraljka Golub, Marcelo Milra. Vol-2021. M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen., Aachen (2016)

8. Dannélls, D., Johansson, T., Björk, L.: Evaluation and refinement of an enhanced ocr process for mass digitisation. In: Proceedings of 4th Conference of the Association Digital Humanities in the Nordic Countries (DHN). CEUR (2019)
9. Dannélls, D., Persson, S.: Supervised post OCR correction of historical Swedish texts: What role does the OCR system play? In: Proceedings of DHN 2020 (2020)
10. Engwall, L., Hedmo, T., Persson, O.: Corpus linguistics in Sweden: Pioneers and their contexts. Kungli vitterhets historie och antikvitets akademien, Stockholm (2019)
11. Fridlund, M., Olsson, L.J., Brodén, D., Borin, L.: Trawling for terrorists: A big data analysis of conceptual meanings and contexts in Swedish newspapers, 1780–1926. In: Proceedings of HistoInformatics 2019. pp. 30–39. CEUR-ws.org, Aachen (2019)
12. Jarlbrink, J., Snickars, P.: Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. Journal of Documentation **73**(6), 1228–1243 (2017)
13. Karsvall, O., Borin, L.: SDHK meets NER: Linking place names with medieval charters and historical maps. In: Proceedings of DHN 2018. pp. 38–50. CEUR-ws.org, Aachen (2018)
14. Rouces, J., Borin, L., Tahmasebi, N.: Political stance analysis using swedish parliamentary data. In: Workshop Proceedings (Vol. 2364). Digital Humanities in the Nordic Countries 4th Conference. CEUR Workshop Proceedings (2019)
15. Rouces, J., Borin, L., Tahmasebi, N., Eide, S.R.: Defining a gold standard for a swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. In: CEUR Workshop Proceedings vol. 2084. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki, Finland, March 7-9, 2018. Edited by Eetu Mäkelä Mikko Tolonen Jouni Tuominen. University of Helsinki, Faculty of Arts, Helsinki (2018)
16. Rouces, J., Tahmasebi, N., Borin, L., Eide, S.R.: SenSALDO: Creating a Sentiment Lexicon for Swedish. In: LREC 2018, Eleventh International Conference on Language Resources and Evaluation. ELRA (2018)
17. Tahmasebi, N., Hagen, N., Brodén, D., Malm, M.: A convergence of methodologies: Notes on data-intensive humanities research. In: Proceedings of DHN 2019. pp. 437–449. CEUR-ws.org, Aachen (2019)
18. Tangherlini, T.R.: The folklore macroscope. Challenges for a computational folkloristics. Western Folklore **72**(1), 7–27 (2013)
19. Viklund, J., Borin, L.: How can big data help us study rhetorical history? In: Linköping Electronic Conference Proceedings, No. 123. Edited by Koenraad De Smedt. Selected Papers from the CLARIN Annual Conference 2015. October 14–16, 2015, Wroclaw, Poland. vol. 123, pp. 79–93 (2016)
20. Waldispühl, M., Dannells, D., Borin, L.: Material philology meets digital onomastic lexicography: The NordiCon database of medieval nordic personal names in continental sources. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 860–867. European Language Resources Association, Marseille, France (2020)
21. Wissik, T., Resch, C.: Researcher Hands-On Training in the Digital Humanities: The ACDH Tool Gallery as an Austrian Case Study. In: Borin, L. (ed.) Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, CLARIN Common Language Resources and Technology Infrastructure. pp. 131–138. Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköping, Sweden (2017)