Collaborating on Language Resource Infrastructures with Non-Research Partners: Practicalities and Challenges

Verena Lyding
Eurac Research, Italy
{verena.lyding,egon.stemle}@eurac.edu

Egon Stemle
Eurac Research, Italy

Alexander König
CLARIN ERIC, The Netherlands
alex@clarin.eu

Abstract

By now, digital infrastructures for language data and tools have become commonplace in the research domain, but their possible benefits are still almost unknown outside of these circles. However, it stands to reason that the data and methods developed there could also be used by non-research language actors like publishing houses or libraries. This article presents a use case within a local language infrastructure project describing our interactions with a newspaper portal that resulted in modern NLP tools being made available via an API to help improve their online search. We describe how this use case was implemented, focusing on the problems that came up, specifically those from the interaction between a research and a non-research institution.

1 Introduction

Large scale research infrastructure projects like CLARIN (De Jong et al., 2018), DARIAH (Edmond et al., 2017) or ELG (Rehm et al., 2021) aim at making language resources and tools available to, sustainable for and easily reusable by their stakeholders. These efforts have proven to create standards and frameworks and have become a reference point for visibility. Yet, up to today, the active involvement of stakeholders and the ambition to attract users to the provided services and tools is challenging. For example, different User Involvement (UI) events of CLARIN helped to provide specific training to a number of research stakeholders[1] and the CLARIN Resource Families initiative (Fišer et al., 2018) links resources of several research stakeholders. Still, stakeholders from industry are hardly found among the users, and experiences from projects that actively involve commercial partners show that the industrial use of the offered services is indeed difficult.[2]

This is related to the naturally slow advancement of large scale, complex and usually abstract projects. In particular, solutions that aim at encompassing various use cases and demands tend to result in powerful yet generic frameworks, as for example, the Component Metadata Infrastructure[3] (Goosen et al., 2014). Those solutions are not always easy to adopt because they require knowledge and technical skills, and it is not always clear from the onset whether the actual use case can be implemented. For this reason, and to bridge the gap between research and application, projects are created that cover domain-specific use cases with the help of large infrastructures (for example, ELEXIS (Woldrich et al., 2021)).

This paper presents a use case from the local language infrastructure project DI-ÖSS[4](Lyding et al., 2019). This project bridges the gap between an existing infrastructure project (CLARIN) and a local community. That is, instead of targeting a specific application *domain* (like e.g. lexicography) DI-ÖSS targets a wider set of *local stakeholders* that are working with language in different ways. By doing so DI-ÖSS aims to connect those local actors to ideas, procedures and solutions from the large infrastructure.

[1] See, for example, https://cmc-corpora2017.eurac.edu/uievent/ and an overview here: https://www.clarin.eu/content/user-involvement-funding#guest-blog-posts

[2] Bleichner et al. (2005) report on a cooperation between two German universities and a part of the archiving division of AIRBUS; Poesio and Magnini (2009) report on a project with data providers of audio, video and text news from Trentino, Italy.

[3] https://www.clarin.eu/cmdi

[4] *Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und –dienste* - Digital infrastructure for the ecosystem of South Tyrolean language data and services

Verena Lyding, Egon Stemle and Alexander König 2022. Collaborating on Language Resource Infrastructures with Non-Research Partners: Practicalities and Challenges. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 88–100. DOI: https://doi.org/10.3384/9789179294441

2 Background

2.1 Local Infrastructure Project DI-ÖSS

The presented use case on an interaction between an online newspaper portal and an NLP service hosted at a research institution was carried out as part of the small local infrastructure project named DI-ÖSS (Lyding et al., 2019) which ran from 2017 to 2021. The aim of the DI-ÖSS project was to connect various types of language actors on the local level to exploit synergies between their activities and goals and the objectives of Eurac Research's Institute for Applied Linguistics (IAL). The IAL is a member of CLARIN-IT and is the initiator and leader of the DI-ÖSS project. The project explicitly aimed at the involvement of non-research partners, which are typically not familiar with infrastructure efforts on the European level. A consortium with four local language actors was established to explore different use cases. Next to the newspaper portal two public cultural institutions, a local library partner and a public culture and language institution, as well as a non-computational research partner working with historical letters were part of the consortium. The model of cooperation between each of the four local language actors and the IAL was an asymmetric project cooperation, with the major workload on the side of the IAL, and a smaller workload on the side of each of the partners, limited to accompanying the use case development with their relevant institutional knowledge. Any active data curation and development work was delegated to subcontractors, which were coordinated by the IAL and paid by the project budget.

2.2 Finding Partners

The first phase of the project consisted of the lead partner at IAL searching for cooperation partners. These partners should be outside of the area of research as the aim of the project was to widen the idea of *infrastructure for language data* beyond the scope of research where it is well established by now. As a first step an extended list was created of institutions that primarily work with language data within the project's dedicated geographical region, the Autonomous Province of Bolzano/Bozen in northern Italy. There can be a case made that any institution is dealing with language data to some degree, but for this list the focus was on institutions where the language data is their main focus. In the end, this list contained about 200 entries that could be grouped roughly into seven categories: archives, libraries, online media, catalogues, cultural institutes, publishing houses and journals. See Figure 1 for their distribution.

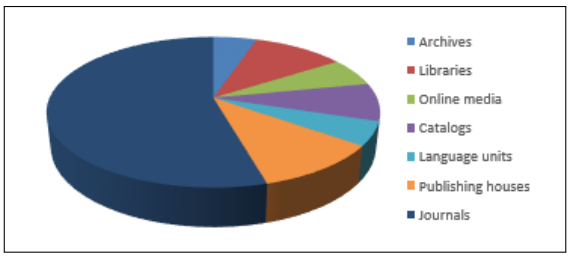

Figure 1: Distribution of institution types in preliminary classification

The overabundance of journals in our long list is due to the fact that they are very visible carriers of local language data, even though a lot of them are not produced by what can be called a "language institution". We kept them in the list, because they can be a rich source of local language data, but drastically reduced their weight in the following steps, namely the interview phase described below.

The IAL did in-depth interviews with eleven individual institutions, trying to cover all of the various categories and also looking at institutions of different sizes. These interviews were used to get some idea of typical processes within these institutions (Lyding et al., 2020) and starting from that develop some possible use cases that could be worked on with such an institution within the scope of the project.

The IAL then invited some institutions as cooperation partners into the DI-ÖSS project. As during the interviews, it was tried to cover a wide range of different institutions. But it turned out to be surprisingly difficult to convince non-research institutions to join this kind of project. We encountered quite strong reservations regarding whether participating in such a project would be worth the institution's time. We assume that this can be explained by the fact that especially institutions organised as a business have to

always calculate whether the time (and therefore money) they invest into such a new project will be met by enough revenue, that is, will they get enough out of it or even more general *what* can they get out of it. For partners from the newspaper publishing world, three out of the four institutions we contacted were not interested in any cooperation even though, apart from the possible technological benefits, a small monetary reward was offered.

2.3 Defining Use Cases

Another problem that paired with this general reservation was the challenging task of envisioning possible use cases that could be explored in such a cooperation. When setting up the cooperation with the newspaper portal, we tried to develop a use case together with the people working there. But we realised that it was difficult for them to see beyond their day-to-day work within an established environment and come up with ideas that could utilise the possibilities lying in such a cooperation with a research institution. This showed that it is difficult for a potential industrial partner to envision a possible use case using NLP tools and other language technology methods because the extent of these methods is not widely known. Therefore, users either have no idea at all what is possible or on the other hand greatly overestimate the power of these tools and come up with ideas that are virtually impossible with today's capabilities. Also, the factor of having to adapt established workflows can pose obstacles for businesses offering professional services. Any adaptation can lead to a possible disruption of a workflow, and it is therefore understandable that non-research partners are particularly wary of committing to 'unnecessary' changes to a running system, even more so if the added value is something abstract like an evaluation metric, a promise of an improved experience, or a functioning prototype with different data or not fully integrated into their usual workflow. While we can envision potential use cases and the expected added value of a project, the cooperation with a research partner and research tools cannot be guaranteed to be as stable and predictable as commercial services. Insofar, it was a fine line we had to walk in order to entice partners with possible opportunities on the one hand, but also not to promise too much.

2.4 Related Work

The research community is in active exchange on language research infrastructure initiatives as conferences by the main players CLARIN, META-SHARE and ELG, and dedicated conference tracks at NLP conferences[5] show. Also, calls for application showcases are widely promoted and fostered with financial incentives[6][7]. Despite this active promotion of the adoption of language research infrastructures by a wider audience, it is extremely difficult to find scientific relations and reports on research - industry cooperations. This even holds for the inter-institutional projects mentioned above.

While it would be very valuable to gain insights on prior experiences with adopting research infrastructure components for use cases from industry, the lack of these types of publications is also comprehensible. Research - industry cooperations are challenging by nature and in the context of project-based initiatives often experimental and small in scale. If achieved, results may remain preliminary and use cases might not always turn out as success stories, thus the motivation to publish about it can be naturally diminished. In addition, scientific conferences generally target substantial scientific contributions and concluded works rather than work-in-progress reports. In conclusion, our search for related works has been without noteworthy results and it remains to the scientific community to encourage more project reporting on less shiny but insightful use cases and cooperations over time.

3 Cooperation with a Newspaper Portal

The use case explored further in this paper is built on the inter-institutional cooperation among the Institute for Applied Linguistics (IAL) at Eurac Research[8] and the local newspaper portal salto.bz[9]. Among the cooperations with the four project partners (see above), we decided to focus on this single cooperation

[5] https://lrec2022.lrec-conf.org/en/calls-papers/2nd-call-papers/
[6] https://www.clarin.eu/content/user-involvement-funding
[7] https://www.european-language-grid.eu/open-calls/
[8] https://www.eurac.edu/linguistics
[9] https://www.salto.bz

as the local newspaper portal was the most commercial/industry partner within the consortium. Moreover, the local newspaper was the only partner we had not worked with before. Accordingly, this collaboration was the most challenging and insightful in terms of identifying a use case and implementing it.

3.1 Cooperation Partner: Salto.bz Newspaper Portal

Salto.bz is a 'news and community portal for South Tyrol'. It has been founded in 2012 as a cooperative society and its news portal is online since 2013.[10] Salto.bz is the first German and Italian bilingual online news portal in the multilingual province of South Tyrol. It was created with the aim to combine journalism and social media communication and offers editorial content of professional salto.bz authors as well as texts, comments and multimedia content provided by its community. It focuses on journalism and information exchange on daily news and analyses on politics, economy, environment and society.

Among the different news publishers in South Tyrol salto.bz stood out by its openness, interest and availability for a cooperation with us as research institution. In contrast to experiences from multiple other attempts to find collaboration partners among local publishers the head office of salto.bz immediately signalled willingness to learn more about the project idea, to discuss specific cooperation possibilities, and to promote the cooperation to its internal workers and to involve them where relevant.

3.2 Use Case: Advanced Search

The initial discussions between the IAL and salto.bz focussed on understanding the structure of the newspaper portal, the uses-cases of their internet users, the types of data which are created and administered by salto.bz and the related workflows from article writing to publishing. The aim was to identify a use case that could benefit from NLP treatment. This way, the IAL could offer an NLP service to salto.bz while, in return, getting access to authentic language data produced in the bilingual context of South Tyrol to carry out linguistic studies.

Analyzing internal workflows at salto.bz, the following things could be observed. Concerning the back-end interaction, it became clear that authors do not perform any language processing activities within the portal. They mainly interact with it to upload and publish new articles. In addition, news writing and publishing is often carried out under time pressure, and any additional activity required before completion is not appreciated by authors, unless the added value is very clear. Concerning the front-end interaction, that is the interaction of readers with the portal, the search functionality within current news and news in the archive showed some shortcomings in terms of search speed, support for multiword searches and search by criteria like publication date, author or ressort.

The use case we jointly identified targets the creation of an improved search service for the articles of the news portal, including recent articles and the entire news archive. To be most useful to salto.bz users, the outline of the improved search service foresaw the following functionalities:

1. Full-text search with support for multiword searches,
2. Facetting by manually created metadata for each article (i.e., author name, publication date and section of the newspaper), and
3. Automatically generated keywords as additional facets to refine the search results.

Thus, the use case implements a service of general interest for the news portal and includes an NLP component prototype which is delivered by the IAL as research partner. The distribution of the work between salto.bz and the IAL of Eurac Research was organised as follows: salto.bz took care of the general technical aspects of the portal programming, while the IAL provided a computational linguistic component for multilingual keyword extraction from news articles. Detailed information on the distribution of work and the interaction between project members across institutions are given in Section 4.3 below.

[10] https://www.salto.bz/de/faq

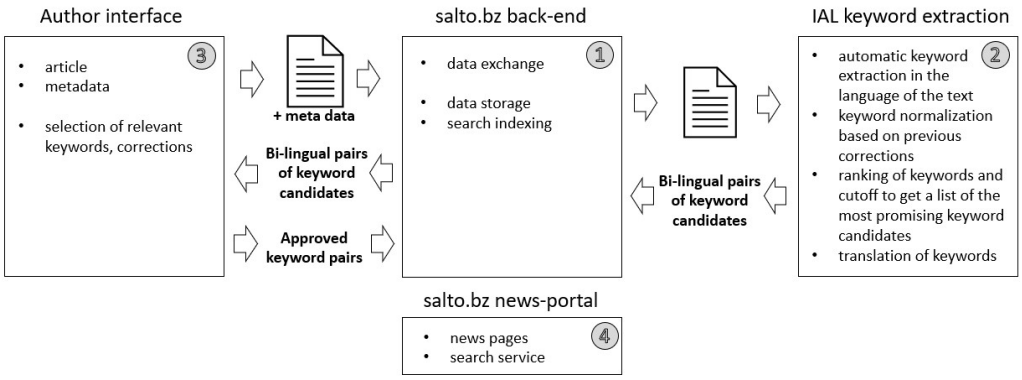

Figure 2: System architecture

3.3 Coordination of the Inter-Institutional Interaction

The inter-institutional cooperation was initiated by the project coordinator of the IAL who contacted the salto.bz head office and proposed a project cooperation. The salto.bz head office agreed to be part of the project and participated in several meetings with the IAL to discuss the working reality and content on the side of the newspaper portal and to identify the specific use case for the cooperation. Once the use case was established, on salto.bz side the overall interaction continued to be handled by the head of the back office, who administered contractual aspects and coordinated the participation of salto.bz employees in the use case, mainly the interaction with the programmer of the portal. On IAL side the cooperation was coordinated by the project lead in consultation with its other researchers. The technical implementation was carried out between the programmer at salto.bz and the project researchers at the IAL. As soon as the system components were in place, also the editor-in-chief of salto.bz got involved on the client side to give feedback on the user interaction within the author's interface. The editor-in-chief also took care of the communication with the authors of salto.bz. Finally, the IAL involved a translation expert to evaluate and curate the automatically extracted keyword pairs.

4 Implementation of the Improved Search Service

4.1 System Design

The extended search service is designed as a distributed architecture with the search interface running on the salto.bz news portal and the computational linguistics text processing being performed at the IAL.

Overall, we can distinguish four components of the system architecture (see also Figure 2):

1. The portal back-end, data storage and search engine (salto.bz[11])
2. The keyword extraction service (IAL)
3. The author web interface (salto.bz)
4. The news search web interface (salto.bz).

The portal is centrally based on Drupal[12], an open-source content management system (CMS) that can be extended with modules to expand its functionality. Thus, the functionality on the salto.bz-portal side was fully integrated into the CMS and its regular workflow. This means the authors enter their articles via a web interface and optionally activate a *Get Tags*-function, which triggers a process on the back-end side

[11] Strictly speaking, the division here is subdivided again: The CPU time, the data storage, the general Drupal and Apache Solr/Lucene installations are provided by an external Internet service provider.

[12] `https://drupal.org`

of the CMS to send text with metadata to the keyword extraction service to retrieve candidate keyword pairs. The keyword extraction service at the IAL receives data from the portal back-end, processes the news article texts and metadata, identifies keywords, sends new keywords for translation to an external service, and returns candidate keyword pairs back to the portal back-end. These candidates are then forwarded to the author's web interface for validation. The author's interface allows to validate, delete or modify candidate keyword pairs before the article is queued for review or publication.

4.1.1 The News Search Web Interface

The news search is based on the integration of Drupal with Apache Solr[13], a popular open source enterprise search platform built on Apache Lucene. The web interface allows for full text search and facetting of results both by metadata information and keywords. Figure 3 shows the results page for the search *Fahrrad (Bicycle)*. The articles that match the query are shown on the right, with their section (e.g. *UMWELT (Environment)*), title, author, publication date, and highlighted text for the query match(es). The left side (or a popover on smaller screens) shows some of the possible facets to restrict the search

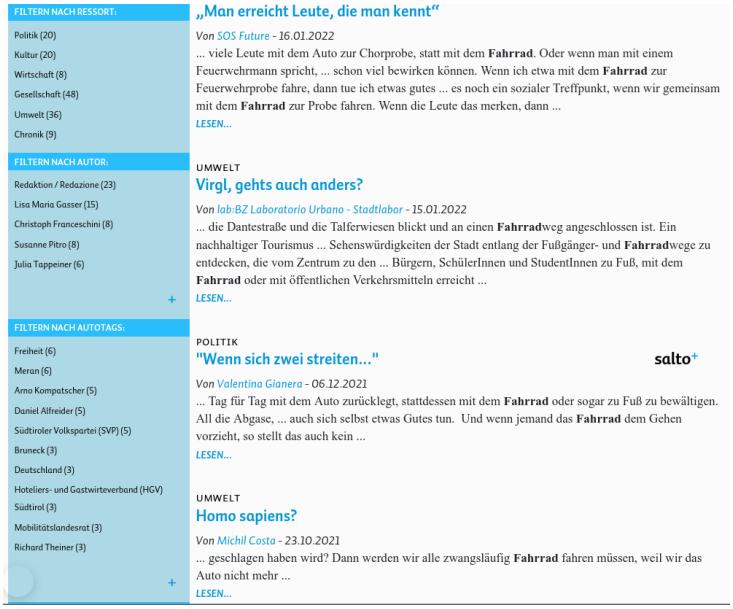

Figure 3: Search results with (some) facets

results: *FILTERN NACH (Filter by) RESSORT (Section)*, *AUTOR (Author)*, and the generated *AUTOTAGS (keywords)*. The first facet, which is at the beginning of the list and not visible any more on this scrolled down view, is *year of publication*. Further down the facets list is also a short info box containing information about search in general and the Autotags feature (our translation):

> A search for several words is possible. For an exact search, the terms can be placed in inverted commas.
>
> The number of hits per category is displayed in the search options. Clicking on a value filters by category.
>
> The category 'Autotags' shows keywords that have been automatically generated using computer linguistic methods. This functionality is the result of a research cooperation with Eurac Research and is currently in beta phase.

4.1.2 The Author Web Interface

The authors of salto.bz enter news content through the author web interface which provides a form with distinct fields for the news body text, title, section, etc. In terms of content creation, news articles and metadata such as section or publishing date are manually created by the authors of salto.bz, while keyword pairs for each text are generated on demand by the keyword extraction service of the IAL. After entering the news text, authors have to actively retrieve and validate keyword pairs through the system. Figure 4 shows the author web interface with the part for the keyword extraction highlighted. Keywords

[13] https://solr.apache.org/

are generated by clicking on the 'Get Tags' button. The automatically extracted keywords will appear below the text field; light grey indicates new keywords and turquoise keywords that already exists in the system. Suggestions can be approved (⊕), which adds them to the text field, or can be cleared from the text field (⊖). New keywords can also be freely added to the text field or any approved one can be changed.

Figure 4: Author's web interface for editing an article and the keyword extraction 'Autotags' highlighted

4.2 Technical Implementation of the Keyword Extraction Service

The keyword extraction tool is a prototype implementation which extracts keyword from an article and uses the Microsoft Bing translation API[14] to translate each extracted keyword into the respective other language. That is, it receives a text in either Italian or German and returns a list of bilingual keyword pairs in the form *German@@Italian*. Together with the developer of the newspaper portal, we defined three requirements for user interaction. The interface should allow (1) to retrieve a list of candidate keyword pairs on demand, (2) to select or deselect candidates according to their relevance and (3) to change or correct the selected candidates if necessary.

Keywords for us are single or multi-word expressions that do not necessarily have to occur in sequence in the text and are extracted by a handful of manually devised rules[15]. Through the rules, we were able to ensure that some peculiarities of typical local naming, for example, names of public administration entities, can be recognised in both German and Italian. The translation was done for each keyword independently of its context - this is a borderline use of the translation service, but for a working prototype we accepted this shortcoming. This implicates that for ambiguous keywords the correct translations cannot be guaranteed by the system, but are enforced in the manual validation step instead.

On the technical side, the exchange between the parties was standardised. All involved parties could develop their system independently, and still, the systems could communicate with each other. To this end, a RESTful application programming interface (REST API) (Fielding, R. T., 2000) was designed and implemented that allows documents to be sent to the IAL for computer-aided linguistic processing, which

[14] https://docs.microsoft.com/en-us/azure/cognitive-services/translator/

[15] See keyword_extractor_salto.py in https://gitlab.inf.unibz.it/commul/di-oss/api-service-salto

can be retrieved after successful processing. The API implements an authentication and authorisation layer that allows us to track which document came from whom so that different processing or licensing agreements can be considered.

The processing of the documents is (optionally) asynchronous in order to be able to take into account long-lasting processing. For this purpose, the API assigns an identification token after successfully transmitting a job, which the remote party must use when returning to query whether processing has already been completed. Once processing is complete, the bilingual keyword pairs suggested by the system can be retrieved and are displayed to authors for further processing as described above.

The keyword pair candidates are compared with those already in the system, and the already known pairs are colour-coded. As explained above, the suggestions must be actively accepted, i.e. there is no mechanism that automatically assigns the suggestions to an article. In addition, the provision of keyword candidates is beneficial but not critical. In the absence of such suggestions, authors can also complete editing an article without automatic suggestions by manually entering their suggestions, which are automatically completed with the suggestions known to the system.

Keyword pairs can also be curated in a separate interface. Changes in this interface are recorded so that traceable changes can be systematically automated. For example, a singular-plural association can be made, which is then recorded as an entry in a file and considered for future proposals. In this way, an author could benefit from both the automated system (the proposal) and the regular maintenance of the taxonomy in a future article. In order to ensure an exchange of information about the acceptance or the content of the accepted keywords, a data reconciliation is carried out at regular intervals. For this purpose, the log file is provided by the developer of the news portal and transferred to our system.

It is important to underline that the keyword extraction service is a prototype implementation not a final product. It has been implemented for the purpose of getting a viable use case up and running within a project cooperation between a research and a business partner. Given this restricted scope and related time constraints, no formal user evaluation (neither with news authors nor with the portal users) has been carried out, but informal feedback on the service has been collected from the news authors throughout the project and indicated that the overall keyword quality was considered acceptable.

4.3 Interaction between Project Members across Institutions

In addition to the overall coordination of the project work across the two institutions (cf. Section 3.3) in particular the implementation work had to be orchestrated between the IAL and salto.bz and the different roles of the project participants. The technical implementation of the different system components were divided as follows: salto.bz was in charge of implementing the back-end of the full text search with facetting by metadata, as well as the user search interface and the integration of the keyword extraction tool within the author's interface. The IAL was in charge of implementing the keyword extraction tool and making it available as an independent service. While the implementation of the components were handled independently by the technical profiles (developer/researcher) at both institutions, several interactions were needed to define technical and user-related requirements. The project core team, composed of researchers at the IAL and the salto.bz developer, regularly met to define which functionalities the search interface and the author interface should include and how they should be presented to users of the portal and newspaper authors. The resulting design specification were then first presented to the salto.bz head office and after to the editor-in-chief to collect feedback for the search and author interfaces. After the details of the system design were set, the core team worked on defining the data exchange formats and protocols, which served as basis for the widely independent implementation of the different components by salto.bz and the IAL. Therefore, once the first version of the entire system architecture was put together again the salto.bz head office and the editor-in-chief were involved for system testing from the users' perspective.

Apart from overall design decisions, the cooperative work on this use case mainly concerned the interaction between the author interface of the news portal and the keyword extraction service offered by the IAL. Both with regards to the technical data flow and the user interaction, decisions had to be taken about the number and order of keyword pair candidates, and about how to select and correct them.

4.4 Required Manual Input of News Authors

The system design, as described above, included the authors' keyword validation and curation activity as a fundamental part of the entire search service. Given that the automatic keyword extraction tool can only work so well in extracting relevant keywords and proposing valid translations for the given context, the project team decided that the manual validation of each keyword pair would be the base condition for including keywords as facets within the public search interface.

Once the new search service with keyword extraction was running in beta-version, all salto.bz authors had to be brought on board. The authors had to be made aware of their task to generate (launch the automated process) and approve (manually select, discard and/or correct) relevant keyword pairs. This required teaching them how to use the tool and motivating them to use it for every article they write. To motivate the authors, the added value of the keyword tool was communicated (the keywords serve as additional search facets), and the editor-in-chief encouraged participation.

The project core team provided guidelines for selecting or correcting keyword candidates. Finally, authors were also instructed to document troubles and errors they encountered, as a functioning feedback loop is essential to maintain and improve the service. In fact, we encountered several situations where authors had noticed an error and stopped using the tool without informing us.

4.5 Updating Existing Data

Given that the assignment of keyword pairs to articles requires human approval, by the time the search interface should go live, we had to find a solution for articles written and published before the keyword extraction tool had been introduced. Asking authors to manually validate keyword pairs for all articles in the archive was not feasible, therefore we had to find a way to update the articles from the archive in an automatic way. We proceeded by applying the automatic keyword tool to all articles of the archive, but kept only those keywords that occurred more than ten times, while discarding all the others. This resulted in a list of more than 2000 keyword pairs, which were manually corrected (e.g. for translation errors) and merged by a translation expert, who was subcontracted for this specific task.

The merging task that is the matching of keyword pairs which are not identical but refer to the same entity or concept was relevant for the keyword list of the articles in the news archive, but is expected to become recurrently relevant also in the future. Given that every day new keyword pairs are generated for new articles, it is likely that variants or semantically similar keyword pairs (almost synonyms) will accumulate and will need to be merged manually to keep the overall amount of keywords under control.

5 Evaluation of the Inter-Institutional Cooperation on Language Infrastructures

The primary goal of the DI-ÖSS project, and more specifically on the cooperation between the IAL and the salto.bz news portal, was to explore as many aspects of an inter-institutional cooperation on language infrastructures as possible. In this respect, the presented use case is to be understood as an all-encompassing feasibility study and not primarily as a technical one. This also means that creating a functioning prototype was one partial and practical aspect to exercise the inter-institutional collaboration on a practical use case. The use case of creating an extended search service, in the first place focused on creating and trial running a workflow that allows for the inter-institutional exchange and processing of texts, their integration into a running newsportal and the interaction of automatically applied procedures (the keyword extraction) with manual processing and validation tasks by the news authors.

In the following subsections, we will present a short evaluation of four relevant aspects of the inter-institutional collaboration: (1) technical interaction and performance, (2) quality of automatically and manually processed data, (3) added value for both institutions, and (4) sustainability of the cooperation.

5.1 Technical Interaction and Performance

On the technical level, the interaction was implemented as a RESTful API without major obstacles, however, a number of factors had to be considered and taken care of: (1) security protocols, (2) monitoring and error messages/reporting, (3) time delay, and (4) updates of services, program versions, etc.

example	error type	correction strategy
Ich bin mit Bozen@@Io sto con Bolzano	entire phrase	limit # of (function) words
Alexander Huber@@Alessandro Huber	translation of person names	list of personal names
Flüchtling@@profugo	differing singular plural conventions	manual correction
Kandidatenliste@@Candidati	translation compounds to MWEs	dictionary look-up

Table 1: Error types of keyword pair candidates

Several interaction steps had to be established for the exchange between the authoring interface of the news portal and the language service provided by our keyword extraction and translation tool.

5.2 Quality of Automatically and Manually Processed Data

The quality of automatically generated candidates of bilingual keyword pairs had to be assessed on two levels: On the one hand the keyword pair has to be formally correct and meaningful in itself, and on the other hand the keyword pair has to make up for a meaningful label of the given text. Throughout development we manually evaluated the generated keyword pairs and identified a number of recurrent error types, such as longer phrases or translations of person names. Table 1 lists common error types and strategies applied for their correction.

On the NLP side, we are aware that keyword extraction and translation are in themselves extensive topics within computational linguistics, but ultimately we opted for pragmatic solutions to get the use case up and running within a restricted time frame. This was also made all the easier by the fact that it was desired from the side of the content creators (the news authors) to have full control over all automatically generated output by being able to check and change the automatically generated keywords as 'suggestions' individually. The quality of the suggestions was one important aspect, but the integration of automatic methods and manual quality control within one workflow was the primary one.

During beta testing the adequacy of keyword pairs for news articles was assessed both by the authors of salto.bz as well as by a translation expert, specifically appointed for the quality evaluation of the bilingual keyword data. While a larger number of problems was identified with the formal correctness of keyword pairs as explained above, fewer issues were encountered with their adequacy to describe the article's context. Indeed, most cases of inadequate keyword pairs related to keywords that fit the text, but are little meaningful as keywords, such as *Datum@@data* (engl. date), *klein@@piccolo* (engl. small), *Michl@@Michl* (a first name) or *Nein@@No* (engl. no).

In addition, we encountered a number of keywords with similar semantics that should eventually be merged into one keyword, (i.e., keywords referring to the same concept such as *offener Brief@@lettera aperta* (engl. open letter) and *Brief@@lettera* (engl. letter) should be merged). Having several keywords for one concept is particularly unfavourable in the context of the portal archive search, for which the generated keywords are used as a search facet and apparent 'duplicates' should be avoided.

Since new keyword pairs get added to new articles over time, the coherence of the overall set of keywords needs checking and merging at regular intervals. As an activity that is asynchronous to keyword assignments to single articles it needs particular attention and detailed knowledge of the database and existing keyword pairs. Therefore, this task is best carried out by a professional with dedicated time for it. Also, the correction of keyword pairs turned out to be difficult for some authors, who might not be fully bilingual. To solve the translation and harmonisation issues, we hired a translation professional that carried out the merging and correction task on a weekly basis during the trial phase.

Since the portal search is publicly accessible by all salto.bz costumers, the keyword results need to live up to a minimum quality standard, which can only be guaranteed with regular curation, and indeed the expert role performing the curation would be needed throughout time, which poses a considerable demand for the sustainability of the use case (see Section 5.4 below).

5.3 Added Value for Both Institutions

As the idea of the DI-ÖSS project was to develop use cases to show the potential synergies of establishing a local small-scale infrastructure with non-research institutions one important measure of success was in how far the work that has been done provided added value for the stakeholders involved. For the IAL the added value is very real. Every time an article is sent to the keyword generator, we save a copy of it to our salto.bz corpus that can be used for linguistic analyses, both of Italian and South Tyrolean German.

For salto.bz there are two groups that have to be looked at separately, the authors working with the new setup and the readers of salto.bz. For the readers the new facetted search provides an obvious added benefit. They can now more easily search through the archive of news articles with the keywords helping them find all the articles on a specific subject.

For the authors, the added benefit is less immediate and only occurs if authors indeed can observe an effect of their extra work of assigning keyword pairs to their texts. By being tagged with keywords articles in the news archive are more likely to be found when readers search for related topics, and might in the future even be used to explicitly link older articles to the newly created ones. In this way, adding the keywords might increase the reach and longevity of an article. At the same time, authors have to (slightly) adapt their workflow by adding and checking the keywords which means additional work for them. Even though the amount is very tiny, authors, especially freelancers, already work under pretty tight deadlines and will try to avoid any delays, even if it is just waiting a couple of seconds for an external service to produce some keywords.

When evaluating the use case with our partners at salto.bz it became clear that already during the official run-time of the project the use of the keyword generator had slowly but steadily decreased.

5.4 Sustainability of the Cooperation

Given that the cooperation started as a project effort of limited duration and with limited resources the question of sustainability is crucial and has to always be kept in mind during the run-time of the project. The project outcome has to provide enough added value for both sides to continue to invest time to keep it running smoothly, while at the same time the use case should be set up in such a way that, once everything is implemented, the amount of maintenance to keep it running is minimal.

As discussed above in 5.3, the added value is significant, but not equally distributed among all parties involved. While for the IAL and the readers of salto.bz the added value is immediate, the added value for the authors - increasing the longevity of their articles - can only be measured once the system has been running for some longer time. This leads to the unfortunate situation that the authors are the ones that have to continuously invest time into keyword validation, while being the ones for which the added value is least obvious at first.

Regarding the amount of maintenance needed for the system, the technical maintenance of the keyword generator should be fairly minimal, mostly consisting of keeping the server running and providing occasional security fixes if necessary. The same should be true for the Drupal UI modifications, though here any general Drupal updates might make also changes to the keyword UI necessary.

Apart from the technical maintenance, one larger maintenance task is the regular curation of keywords. To ensure the quality of the service in the long term, the news portal personnel has to regularly merge and check newly introduced keywords (see Section 5.2 above). We assume that the set of keywords will naturally consolidate over time to a certain extent, which means that the amount of work involved in merging and correcting will continuously decrease, but never reach zero.

A possible solution we have discussed to reduce this maintenance workload is to move to a semi-static model: A closed set of keywords is set at a given time, and only keywords from this closed set are assigned to new articles. This closed set is occasionally updated to include the most relevant new keywords (”hot topics”) that occur over a more extended time.

As described above, the adoption of the system by the salto.bz authors decreased significantly during the project’s run-time and we assume it would need new incentives to keep their engagement up. We therefore have decided to move the system into a more automatic state, where keywords from a closed list will be assigned automatically to new texts, removing the added work from the author while still

maintaining a system that assigns keywords to all new texts. In the mid term it is foreseen to revisit the implementation of the keyword extraction tool with the aim to provide an even more reliable service, which would naturally reduce the amount of manual curation needed.

6 Conclusions

The reported work on the use case helped better understand what is needed to establish infrastructure cooperations with non-research partners and what are particular challenges.

The most noteworthy challenge we encountered relates to establishing a cooperation and defining common use cases. We observed that besides a lack of awareness about ongoing language infrastructure initiatives, understanding what it can bring in terms of added value is missing on the side of non-research language actors. Resolving this issue requires extended interdisciplinary communication efforts to identify real needs of business partners and map them to existing LTI solutions. In order to start from tangible scenarios it would be desirable that infrastructure initiatives like CLARIN worked towards a portfolio of use cases for cooperations with non-research stakeholder groups. The recently started ENRIITC project[16](McEntee, 2022) looks like it might be making steps in this direction, with its plan of establishing a network of Industrial Liaison and Contact Officers to facilitate cooperation between research and industry institutions. It also turned out to be highly relevant to reach a common understanding of the objectives and expectations in relation to the use case, as well as a detailed analysis of the workflow and roles needed to implement it right from the start of the project in order to create a sustainable initiative. In the specific use-case of integrating a keyword extraction service into a news portal we ended up with a workflow that was driven by several layers of indirection. While design decisions were taken by the project coordination and technical staff, the immediate contribution to make the keyword assignment happen was required from the news authors, and the results would benefit mainly the news readers. The missing alignment of expectations of all involved parties resulted in a reduced commitment from the side of the authors and thus undermined the sustainability of the effort (see 5.3 and 5.4 for some more details).

A second difficulty relates to integrating the technical implementation with established workflows of a partner institution and fostering the adoption of new procedures. Overall, the communication and decision-making processes required interactions well beyond the technical level and concerned management and editorial participation to a considerable extent. Also, because the realisation of the use case impacted the customer-facing newspaper portal search interface, many people outside our direct contacts paid great attention to all the changes. This experience shows that the workflow planning is a highly complex problem in itself even before detailed aspects of technical solutions or performance of its individual components come in. Again this shows that a highly interactive communication effort between all partners is inevitable and an asynchronous project cooperation with the major workforce on the side of the research partner is not realistically feasible. For future endeavours this suggests that truly interdisciplinary and inter-institutional project cooperations should be targeted, as mentioned above.

Finally, to create a sustainable service addressing questions of quality control and long term maintenance of the service become crucial. The fact that through this project, the news portal created a dependency on an external service as part of their daily workflows underlines the importance of sustainability of infrastructure services and strategies for long term maintenance, which both pose unresolved challenges, not only in this kind of interdisciplinary cooperation but in the whole field of technical infrastructure.

We conclude that up until today, establishing any infrastructure cooperations with non-research partners requires substantial efforts in communication, workflow planning and technical solution building on both ends, the research partner and the industry partner. As long as no portfolio of use cases and applications exist that can be re-purposed, successful cooperations can likely only be created in the context of bigger funded projects. They would need to bring together research and industry partners for several years of intense collaboration, since as of today, on the side of the non-research client of European infrastructures, both awareness for what is doable and support for implementation is still greatly lacking, while on the research side, technical solutions are often not at the level of being ready for the market without considerable customization.

[16] https://enriitc.eu/

Acknowledgements

We would like to thank Monica Pretti for the detailed manual analysis and correction of the first comprehensive sample of automatically generated keyword pairs and the harmonisation of newly added keyword pairs during a limited testing period.

References

Bleichner, M., Giesbrecht, E., Gust, H., Leicht, E.-M., Ludewig, P., Möller, S., Müller, W., Schmidt, M., Stefaner, M., Stemle, E., and Wilke, K. 2005. *ASADO: The Analysis and Structuring of Aviation Documents - Final Report*, Institute of Cognitive Science at the University of Osnabrück and Institute of Applied Linguistics at the University of Hildesheim. https://api.zotero.org/users/332053/publications/items/KSJ9ECLV/file/view.

de Jong, F. M. G., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, Dieter 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan, European Language Resources Association (ELRA). http://dspace.library.uu.nl/handle/1874/364776.

Edmond, J., Fischer, F., Mertens, M., and Romary, L. 2017. The DARIAH ERIC: Redefining research infrastructure for the arts and humanities in the digital age, *ERCIM News*, 111.

Fielding, R. T. 2000. REST: Architectural styles and the design of network-based software architectures, *PhD thesis*, University of California, Irvine http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN's Key Resource Families, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan, European Language Resources Association (ELRA). https://aclanthology.org/L18-1210.

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Ďurčo, M., and Schonefeld, O. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure, *Selected Papers from the CLARIN 2014 Conference*, p. 36–53.

Lyding, V., König, A., Gorgaini, E., Nicolas, L., and Pretti, M. 2019. DI-ÖSS - Building a digital infrastructure in South Tyrol, *Selected papers from the CLARIN Annual Conference 2018*, Pisa, 8-10 October 2018 / edited by Inguna Skadina, Maria Eskevich, Linköping Electronic Conference Proceedings, 159(10), p. 92–102.

Lyding, V., König, A., and Pretti, M. 2020. Digital Language Infrastructures–Documenting Language Actors, *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 3457–3462.

McEntee, J. 2022. Building bridges between big science and industry, *Physics World*, 35(1), IOP Publishing, p. 8–9i, https://doi.org/10.1088/2058-7058/35/01/10.

Poesio, M. and Magnini, B. 2009. Content Extraction Meets the Social Web in the LiveMemories Project, *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009 (AT4DL 2009)*, Bozen Bolzano University Press, p. 42–45.

Rehm, G., Piperidis, S., Bontcheva, K., Hajic, J., Arranz, V., Vasiļjevs, A., Backfried, G., Gómez-Pérez, J., Germann, U., Calizzano, R., and others 2021. European Language Grid: A Joint Platform for the European Language Technology Community, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 221–230.

Woldrich, A., Goli, T., Kosem, I., Matuška, O., and Wissik, T., 2021. ELEXIS: Technical and social infrastructure for lexicography, *K Lexical News*, (28), Zenodo, p. 45–52,https://doi.org/10.5281/zenodo.4607957