Introducing Maalr: A Modern Approach to Aggregate Lexical Resources

Claes Neuefeind and Stephan Schwiebert

University of Cologne, Department of Linguistics, Linguistic Data Processing
Albertus-Magnus-Platz, 50923 Köln, Germany
E-mail: {neuefeind,schwiebert}@spinfo.uni-koeln.de
http://www.spinfo.uni-koeln.de

Abstract. In this paper, we introduce Maalr, an open source framework for collaboratively building web-based lexical resources.[1] Exclusively using freely available technologies, Maalr is dedicated to enable cost effective dictionary projects. The paper describes the basic concepts behind Maalr and gives some details of the technical implementation.

Keywords: Lexica, Collaborative Editing, Online Dictionaries

1 Introduction

Online dictionaries are an inherent part of the web. As the web itself, these resources are highly heterogenous in their appearance and conceptual approach, ranging from dictionaries implementing a straightforward lemma mapping to usage based lexica backed up by corpora. Many of these resources allow for user interaction, e.g. to discuss language-specific questions or most notably to ask for or to suggest new entries. By consequently adopting the latter feature, the Maalr framework is meant to be used to build lexical resources 'from scratch', without the need of having pre-compiled sets of data.

However, in this paper we focus on a practical application of the Maalr framework to the Romansh language, namely the *Pledari Grond Online* (PG)[2]. Initiated in 1982 by the *Lia Rumantscha*[3], the PG still is a consequently growing resource which is regularly updated in an offline procedure: After collecting user suggestions via the web page and by email, new entries are edited by experts and written directly into the database. By using Maalr, the process of modifying and expanding the lexical database can in future be done collaboratively involving the language community.

[1] http://github.com/spinfo/maalr

[2] http://www.pledarigrondonline.ch. The PG is the central lexical resource for Rumantsch Grischun, the standardised official language of the Canton Graubünden, which was realised according to proposals of Heinrich Schmid (see [2]).

[3] http://www.liarumantscha.ch

2 Introducing Maalr

First and foremost, Maalr[4] is meant to be used as an online dictionary to search for a translation. However, the integration with the community has at least the same relevance: aiming at collaboratively building lexical resources, the exchange between speakers and professional linguists is simplified as much as possible.

In Maalr each lemma can be modified instantly via the search interface, without requiring a login. Additionally, visitors can ask for new entries, optionally suggest a translation, enter a comment, or leave an email address to receive a reply. Although anonymous and unverified, all new entries and modifications will be stored and indexed, initially being marked as 'unverified'. For each database entry, Maalr keeps a version history, and whereas by default only the most recent approved version is displayed, visitors can enable a query parameter to additionally list unverified entries or entry versions. Expert users, however, are provided with a dedicated editor for verifying or rejecting new entries. This editor (shown in figure 1) offers multiple filtering methods, to show all modifications done within a specific time, by a specific user or from a specific IP-address.

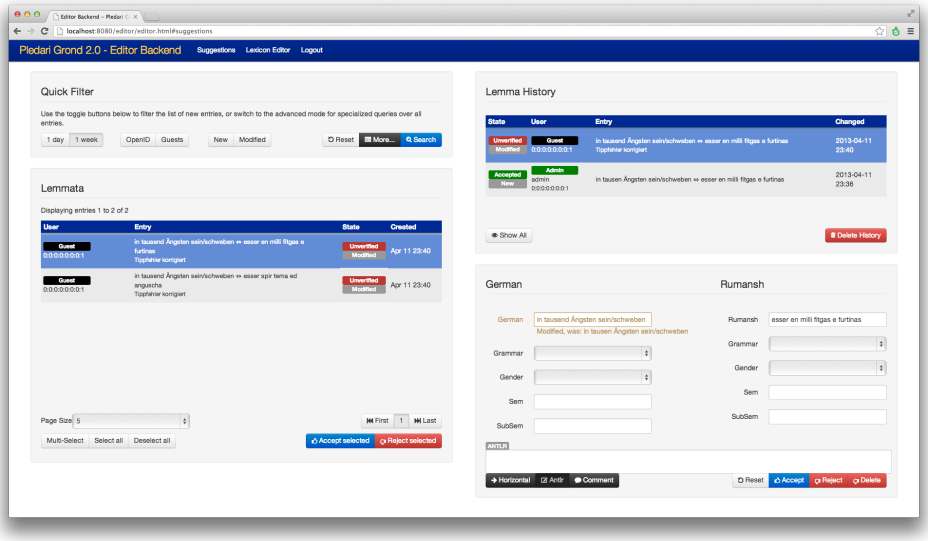

Fig. 1. Maalr's editor interface for approving and rejecting new entries suggested by the community. Lemmas are edited via a form (down right). The editable fields are a subset of the database entries and can be externally configured for separate use cases.

[4] During our project planning, we stumbled across the Wikipedia entry of Josua Maaler, a Swiss lexicographer who created one of the first dictionaries for the German language in 1561, and we decided to give our software his name, in a Web 2.0 version. Nevertheless, we hope that Maalr will be more useful than Maaler's "Die Teütsch spraach", which has never been reprinted or republished.

To handle the different user roles, Maalr facilitates a basic role system. Users can log-in to the platform by choosing a login mechanism such as *OpenID*[5], *OAuth*[6], or *Mozilla Persona*[7]. When logged in as administrator, different rights can be assigned to the accounts. Table 1 gives an overview of the available roles.

Role	Modify, Suggest	Accept, Discard	Verification needed	Administrate DB & Users
Guest	+	–	+	–
External	+	–	+	–
Expert	+	+	–	–
Admin	–	–	–	+

Table 1. The Maalr role system: The default role is 'Guest', changing to 'External' at first login, suggestions and modifications now being marked with the given user details. 'Expert' can only be assigned by the 'Admin' role.

3 Implementation Details

We use a NoSQL-database, namely the document-oriented *MongoDB*[8] storage system as backend (see figure 2), mainly for two reasons: First, Maalr was designed to be open to different languages as well as to conceptual changes or improvements in the lexical micro structure, which is defined via an external configuration file: With a document-oriented database it is possible to add new attributes to a lemma, without the need to modify the whole database schema. Second, when thinking of lemmas as separated and versioned 'documents', this concept is very close to the one realized in the search engine API *Lucene*[9], which is used to manage the visitor queries in the system. Due to this conceptual similarity, we could implement a simple but robust 'mirroring' technique to create and maintain a high performance search index of all lemmas in the database.

The middleware is implemented with *Spring*[10], most notably *Spring MVC*, which expose services and controllers. Besides, it provides simple but well-designed ways to render the data in other formats. Currently Maalr provides a basic export to XML for backup purposes, optionally anonymized (containing the lemma information only) or as full export of all database entries (including the full history and user information). Furthermore, the data is rendered into JSON[11], which

[5] http://openid.net/
[6] http://oauth.net/
[7] http://www.mozilla.org/en-US/persona/
[8] http://www.mongodb.org/
[9] http://lucene.apache.org/
[10] http://www.springsource.org/
[11] http://www.json.org/

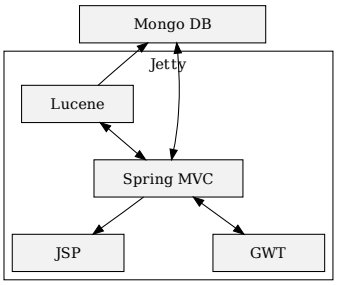

Fig. 2. The Maalr system architecture

allows to easily integrate the data into new applications, such as spell checkers or optimized 'apps' developed for mobile devices. Beyond that, other formats like TEI-XML (*Text Encoding Initiative*[12], see [3]) or LMF (*Lexical Markup Framework*, see [1]) can easily be integrated, both for data export and import.

On client-side we use two different technologies: First, SpringMVC and *Java Server Pages* (JSP) to support 'traditional' browsers as well as search engine crawlers. Second, to simplify the usage of the framework and to allow visitors to interact with it, we used the *Google Web Toolkit* (GWT)[13] to add Ajax-Features to some relevant parts as, for instance, the query interface or the user dialogs.

4 Conclusion

The open source framework Maalr introduced in this paper enables the collaborative construction of web-based lexical resources. Maalr combines a high-performance search engine with a highly responsive, intuitive user interface that simplifies the exchange between the community and professional linguists.

Although in this paper we described Maalr in the light of a concrete adaption to the Romansh dictionary *Pledari Grond Online*, Maalr can easily be adapted to virtually any other language without the need of having pre-compiled sets of data as a starting point. Maalr thus aims at enabling specialized lexicon projects where no or only little data is available, e.g. in the case of minority languages.

All software is being developed openly under http://github.com/spinfo/maalr.

References

1. Francopoulo, G. (ed.): LMF - Lexical Markup Framework. ISTE / Wiley (2013).
2. Schmid, H.: Richtlinien für die Gestaltung einer gesamtbündnerromanischen Schriftsprache Rumantsch Grischun. Chur: Societad Retorumantscha (1982).
3. TEI Consortium (eds.): TEI P5: Guidelines for Electronic Text Encoding and Interchange [Version 2.3.0]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (07.06.2013).

[12] http://www.tei-c.org/
[13] https://developers.google.com/web-toolkit/