A Web Tool for Building Parallel Corpora of Spoken and Sign Languages

Alex Becker[*], **Fabio Kepler**[*†], **Sara Candeias**[‡]
[*]UNIPAMPA - Federal University of Pampa – Alegrete/Brazil
[†]L2F/INESC-ID – Lisbon/Portugal
[‡]Microsoft LDC – Lisbon/Portugal
[*]alex@porthal.com.br, [†]fabio@kepler.pro.br, [‡]t-sacand@microsoft.com

Abstract

The main objective of this work is to build an online tool for manually annotating texts in any spoken language with SignWriting in any sign language. The existence of such tool will allow the creation of parallel corpora between spoken and sign languages that can be used to bootstrap the creation of efficient tools for the Deaf community. As an example, a parallel corpus between English and American Sign Language could be used for training Machine Learning models for automatic translation between the two languages. Clearly, this kind of tool must be designed in a way that it eases the task of human annotators, not only by being easy to use, but also by giving smart suggestions as the annotation progresses, in order to save time and effort. The tool was implemented in the Java Web platform using the JSF framework (Java Server Faces) and an MVC architecture (Model-View-Controller). Figure 1 shows the application domain diagram. Figure 2 shows the process of creating and annotating a document. It starts with the user selecting or creating a corpus. After that, the user creates a raw document (in Portuguese, for example) and runs the document preparation process, in which the system segments the document into sentences and then into tokens. Raw documents may be either manually uploaded as text or automatically fetched from Wikipedia given their URL. The user can then start selecting sentences to annotate. The system searches for candidate signs for each word (or phrase) and ranks them according to similarity, usage frequency, and in the near future by context. The user then combines or separates words and selects the best sign for each entry. As it is common in sign languages, some words in spoken languages have no sign. These are just left empty by the user. Once every sentence is annotated, the document itself is marked as annotated. Figure 3 shows a screenshot of the sentence annotation interface (with a sentence in Portuguese). User selected signs are shadowed and highlighted in green, and different operations are available both for words and signs, like concatenating/splitting and adding a new FSW, respectively. To add a new FSW (Slevinski, 2015) the user can just write or paste its string, but to ease the task, she can also draw the sign using the SignMaker tool (Jr, 2015), which was embedded in the system as shown in Figure 4. This tool allows the user to draw a sign by selecting its composing symbols, which are organized by the sets mentioned earlier: handshapes, facial expressions, body locations, orientation, contact, and movement. It automatically generates the corresponding FSW, which can then be used in the sentence annotation. Finally, an important functionality and the tool's main reason to be is that, at any moment, the user is able to export a corpus or a collection of documents in order to generate a parallel corpus in a simple file format, which can then be used in other tools. By building a collaborative, online, easy to use annotation tool for building parallel corpora between spoken and sign languages we aim at helping the development of proper resources for sign languages that can then be used in state-of-the-art models currently used in tools for spoken languages. There are several issues and difficulties in creating this kind of resource, and our presented tool already deals with some of them, like adequate text representation of a sign and many to many alignments between words and signs. Our next step is to improve the searching and ranking of candidate signs by considering word inflections and by building language models for sign sentences. Reporting the current usage status, we are starting arrangements with professors and researchers at UNIPAMPA for an initial effort in annotating a Portuguese-LIBRAS corpus. Finally, and more importantly, help from the community in developing or using the tool would be very well welcomed.

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

Keywords: Sign language recognition, corpora creation, crowd-sourcing

1. References

Jr, S. E. S. (2015). Signmaker tool.

Slevinski, S. (2015). The signpuddle standard for signwriting text.

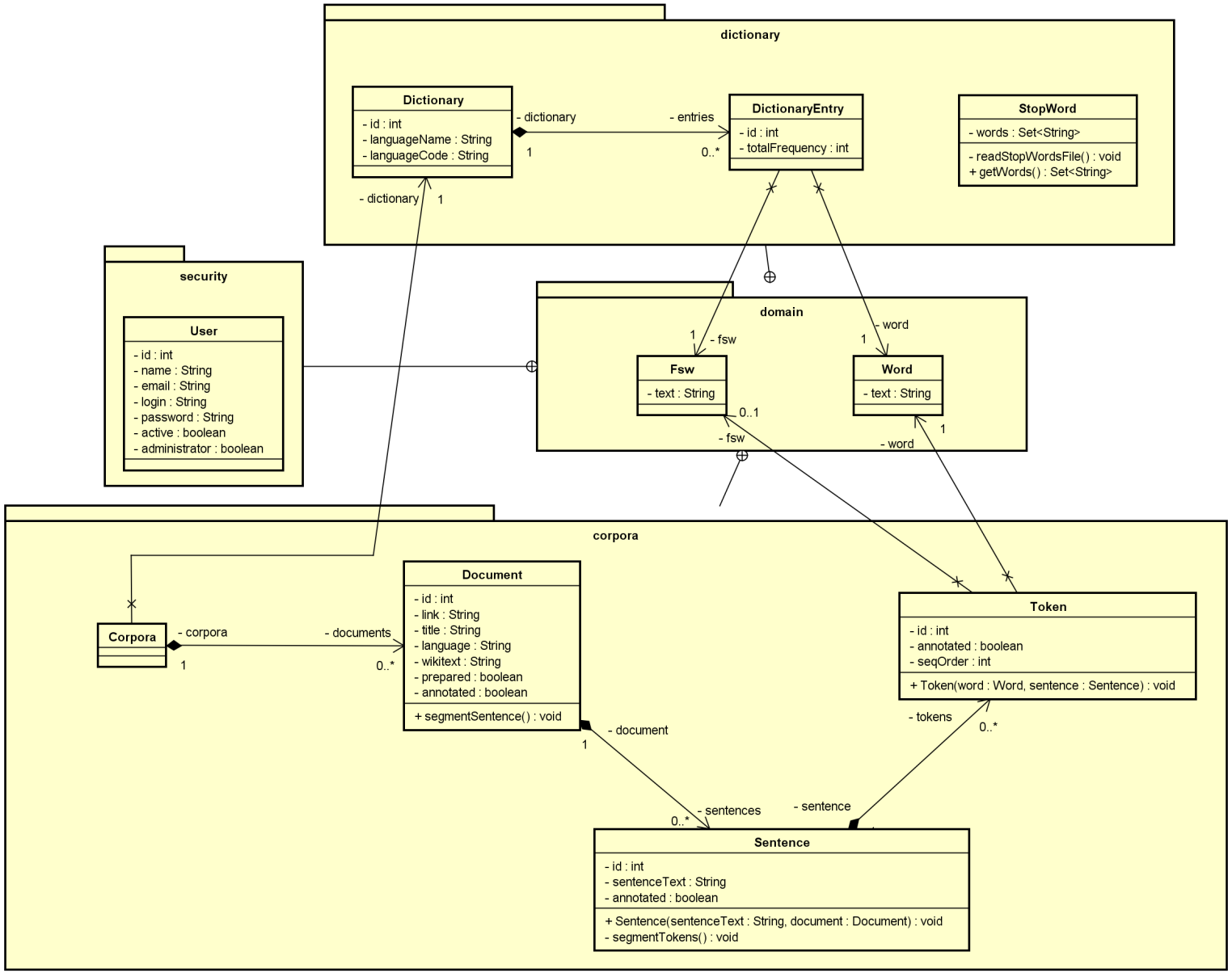

Figure 1: Domain diagram.

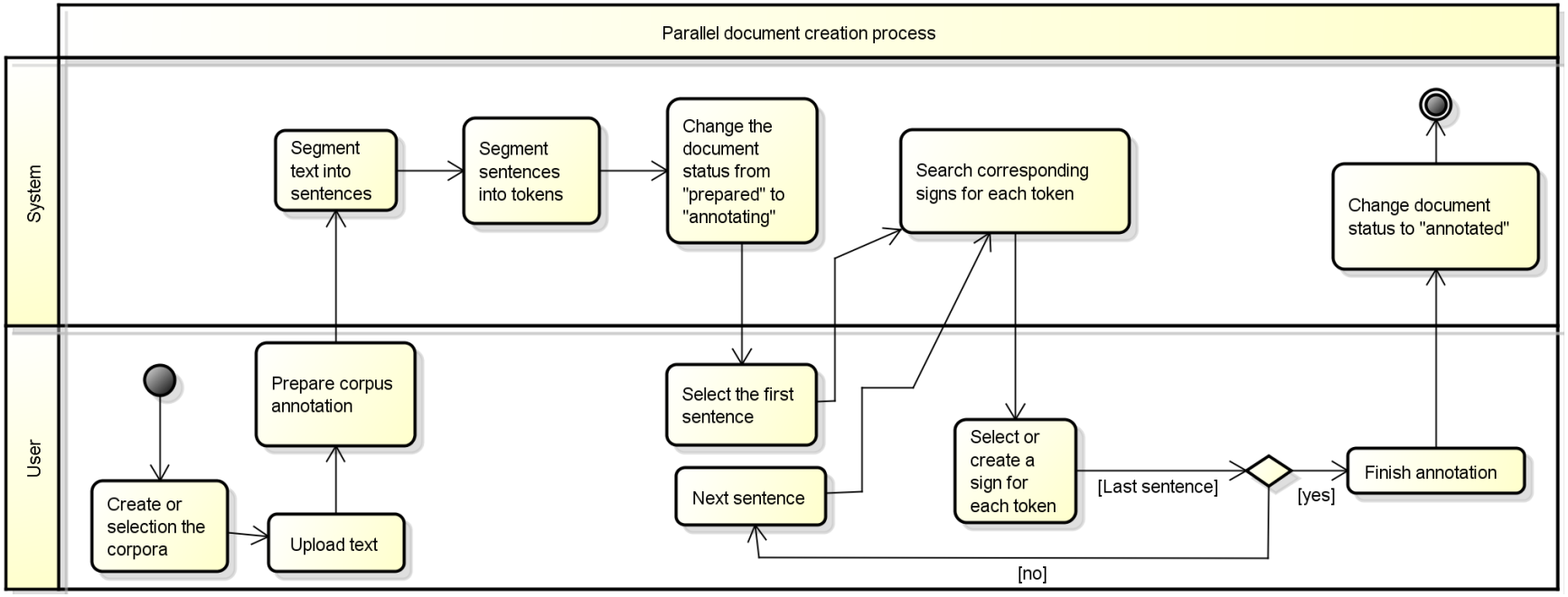

Figure 2: Process for creating a parallel document.

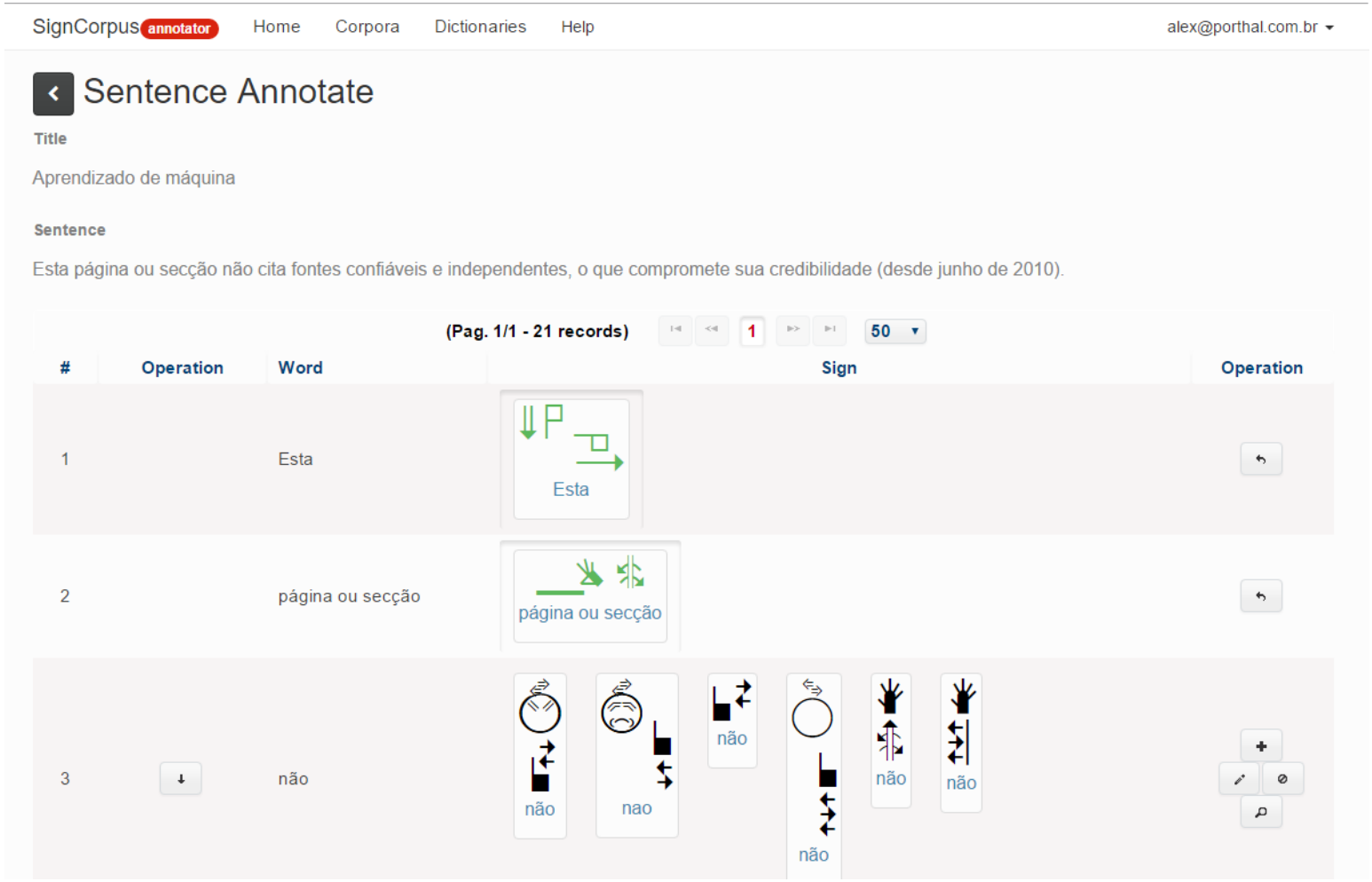

Figure 3: Screenshot of the sentence annotation interface.

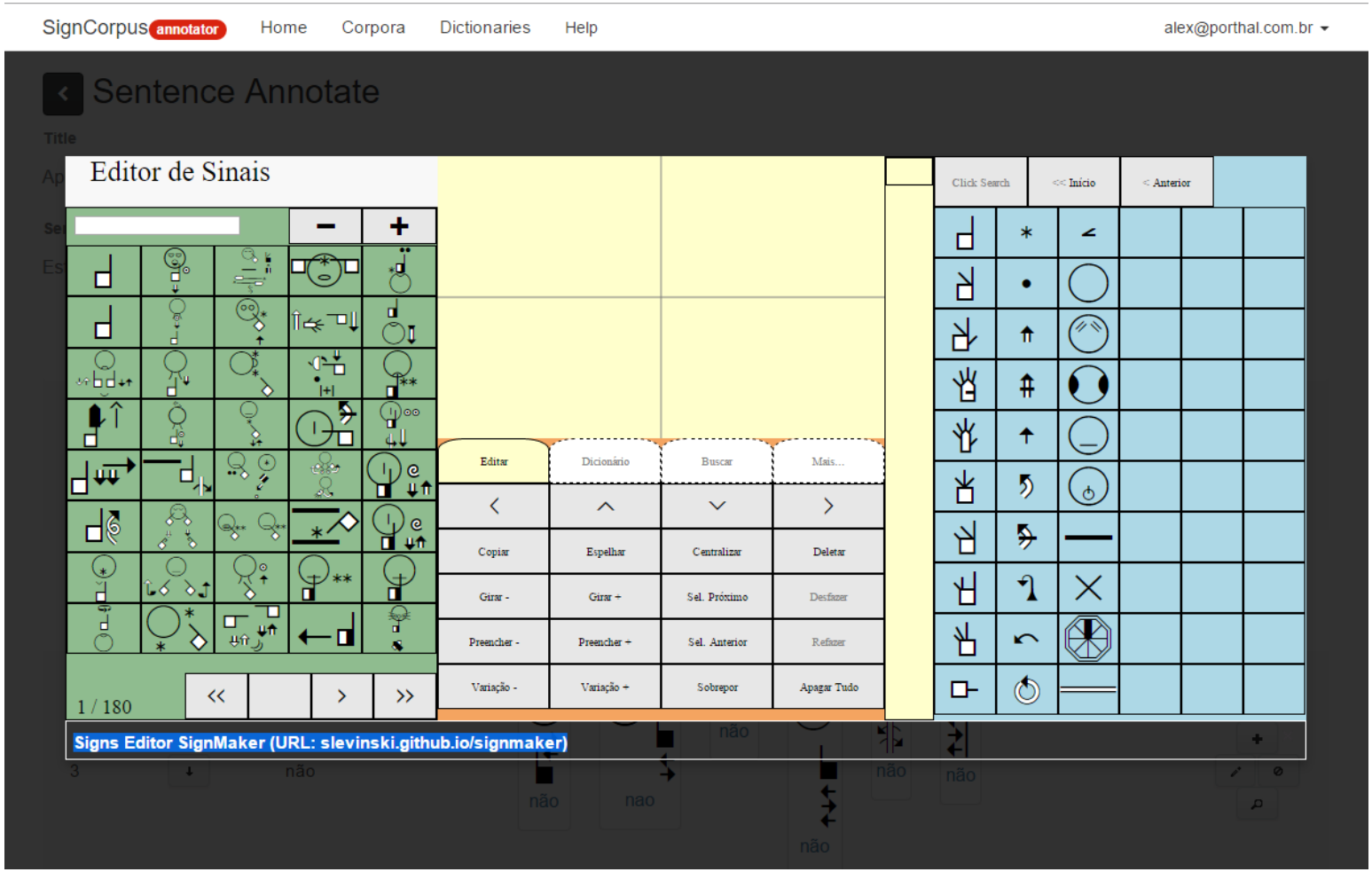

Figure 4: Screenshot of the SignMaker editor interface.