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

ALEX MALMANN BECKER

FÁBIO NATANAEL KEPLER

SARA CANDEIAS

Authors

Software Engineer
Master's degree by UFSCar
CEO at Porthal Sistemas
Soledade, RS, Brazil
Education: UNIPAMPA

Business Development Manager at Microsoft
Lisbon, Portugal
Previously:
INESC-ID, Instituto de Telecomunicações,
Fundação para a Ciência e a Tecnologia (FCT),
University of Aveiro

Visiting researcher
L2F/INESC-ID, Lisbon, Portugal
Professor
UNIPAMPA, Alegrete, Brazil

Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

SignCorpus Annotator

Final Remarks and Future Work

History and Context

Southern Brazil

Rio Grande do Sul

- At the extreme south
- Borders Uruguay and Argentina

Southernmost half lies inside "the Pampas"

- Lowlands that cover 750k km2 and extend further into Uruguay and Argentina

History and Context

UNIPAMPA

Federal University of Pampa

- First activities on October, 2006
- Officially created on January, 2008
- 10 campuses across the Pampas

History and Context

UNIPAMPA

As of last year:

- 64 undergraduate courses
- 27 specializing programs
- 11 masters programs
- 2 PhD programs

Personnel:

- 10,935 undergrad students
- 1,251 graduate students
- 803 professors
  - **10 Deaf professors**
- 835 technical staff
- 375 outsourced staff

History and Context

UNIPAMPA - Alegrete

7 undergraduate courses (Software Engineering, …)

2 master's degrees

1,500 students

90 professors

- 70% with PhD
- **1 Deaf professor**
- **1 sign language interpreter**

89 staff

46 ha total area

8,700 m2 built

History and Context

Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

SignCorpus Annotator

Final Remarks and Future Work

Introduction

Over 200 distinct sign languages in the world.

70 million deaf people over the world.

5.7 million people with hearing impairment in Brazil.

Children who lose hearing before beginning to speak have a sign language as their native language.

Among several proposals for writing sign languages, the most prominently is the SignWriting.

The SignWriting system defines sets of symbols for handshapes, facial expressions, body locations, orientation, contact, and movement.

Introduction

Objectives:

- To build an online tool for manual annotation of texts in any spoken language with SignWriting in any sign language.
- To allow the creation of parallel corpora between spoken and sign languages.
- To design it in a way that it eases the task of human annotators by giving smart suggestions as the annotation progresses.
- A parallel corpus between English and American Sign Language could be used for training Machine Learning models for automatic translation between the two languages.

Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

SignCorpus Annotator

Final Remarks and Future Work

Sign language

Sign languages are the main way of communication in the Deaf community and with the listening population.

It’s not considered a universal language :

- Brazil – LIBRAS (Brazilian Sign Language)
- Portugal – LGP (Portuguese Sign Language)
- EUA – ASL (American Sign Language)

It has differences from one country to another or even from region to region, depending on each culture.

LIBRAS - Second Official Language of Brazil.

SignWriting Representation

Signs stored as images have limited applicability.

Formal SignWriting (FSW) is the latest format for encoding signs.

FSW encodes logographic words (signs) as strings.

M518x517S16d10494x467S33e00482x482S31b00482x482S21900496x456S20500475x476

Parallel corpora

It’s a set of texts where tokens (words) are aligned between a source language and target language.

Portuguese <-> LIBRAS (SignWriting)

Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

SignCorpus Annotator

Final Remarks and Future Work

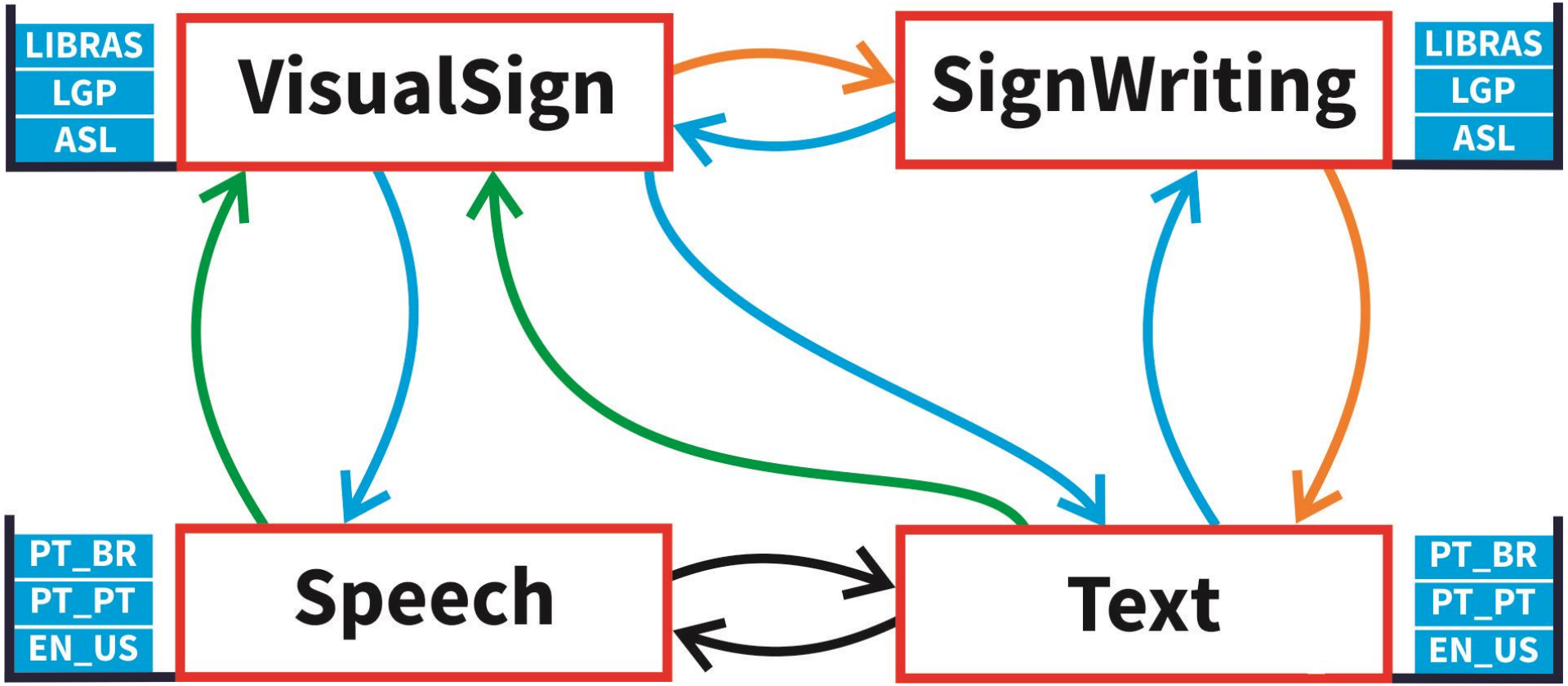
Related Work

We could not find a specific tool for creating parallel corpora in SignWriting.

SignPuddle Online:

- It has a dictionary in Portuguese – LIBRAS (SignWriting).
- Perform simple translation from the dictionary, generating the FSW.
- It could be used to create a parallel corpus, however:
  - Annotation process time consuming and inflexible.
  - External tools needed to perform the entire process annotation.

Related Work



Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

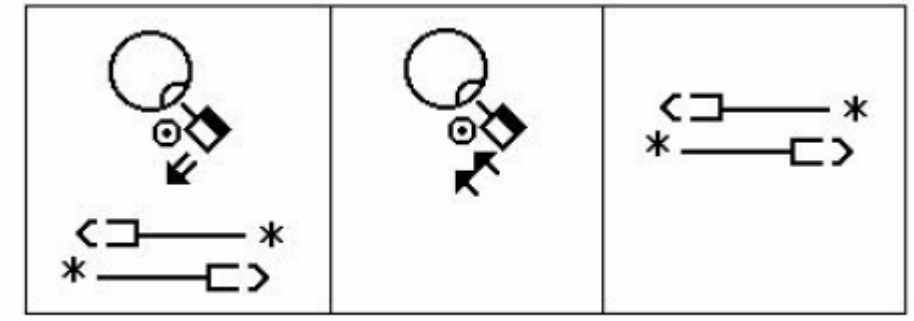
SignCorpus Annotator

Final Remarks and Future Work

SignCorpus Annotator

Problems and Difficulties:

- **One sign to many words:**
  - Sign languages have limited or none at all:
    - Determiners, prepositions, conjunctions, verb conjugations.
  - Also have compound nouns



Bebê do sexo feminino | Mulher | Bebê

SignCorpus Annotator

Problems and Difficulties:

- **Many signs to one word:**
  - Spelling Normalization:
    - “With SignWriting it is possible to have several strings have the same exact 2-dimensional visual appearance.”
    - “It is unlikely that two writers will produce the axact same spelling for any sign.”

SignCorpus Annotator

Current Resources:

- SW icon server: https://github.com/Slevinski/swis
- Javascript library: http://slevinski.github.io/sw10js/
- True Type Font: iswa.ttf
- API and other resources: http://swis.wmflabs.org/

SignCorpus Annotator

Challenge:

- Develop an easy to use tool.
- Perhaps the present form is not the best.
  - Alignments are problematic.

SignCorpus Annotator

Uses SignWriting and an existing tool for constructing new signs.

- Integration SignMaker Signal Editor.

Supports multiple sign and spoken languages.

Allows collaborative annotation.

Provides annotation suggestions based on previous annotations.

Supports importing an initial dictionary from the SignPuddle portal.

Document Import Wikipedia from the URL.

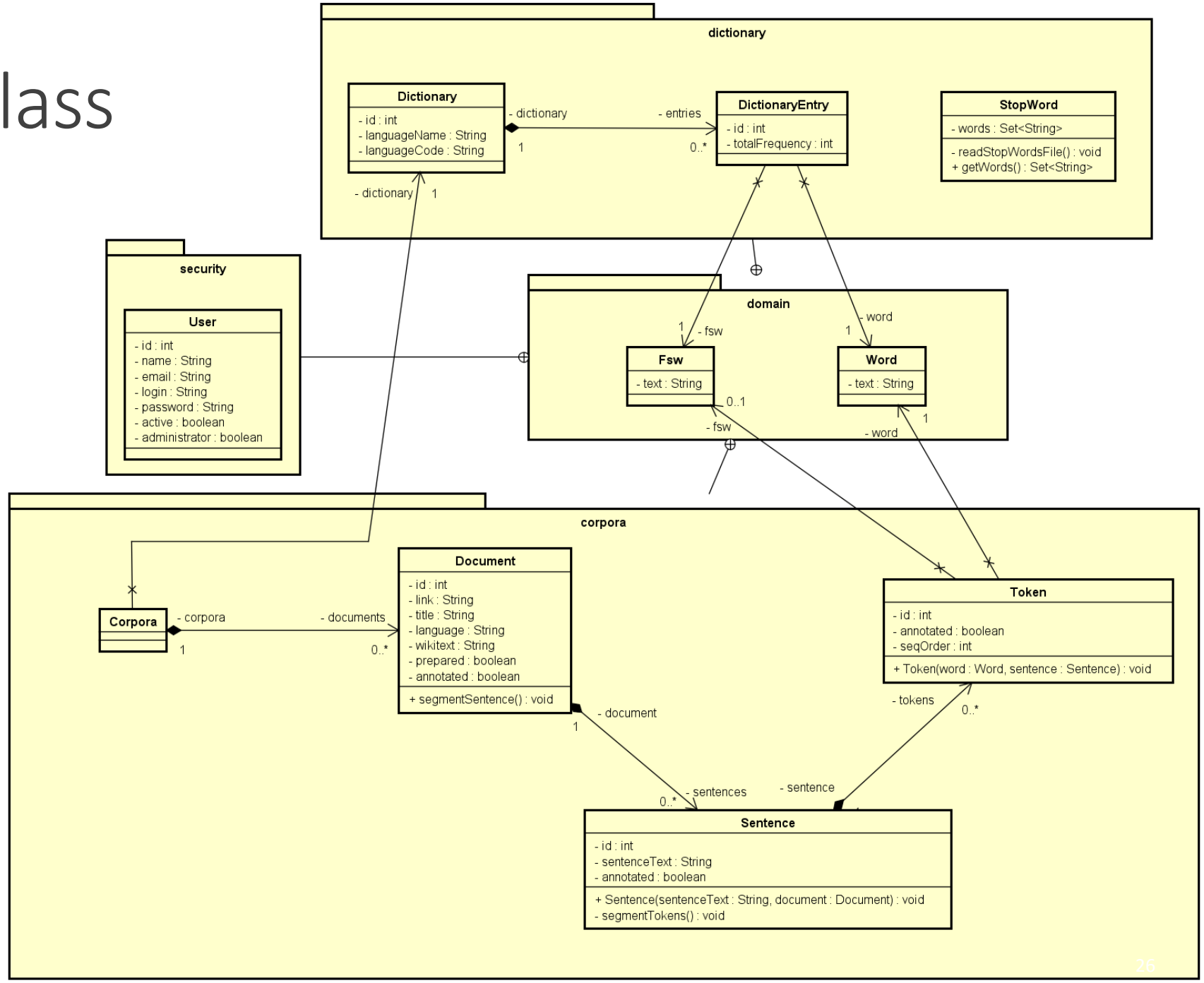
Export corpus Parallel in a txt format.

SignCorpus Annotator

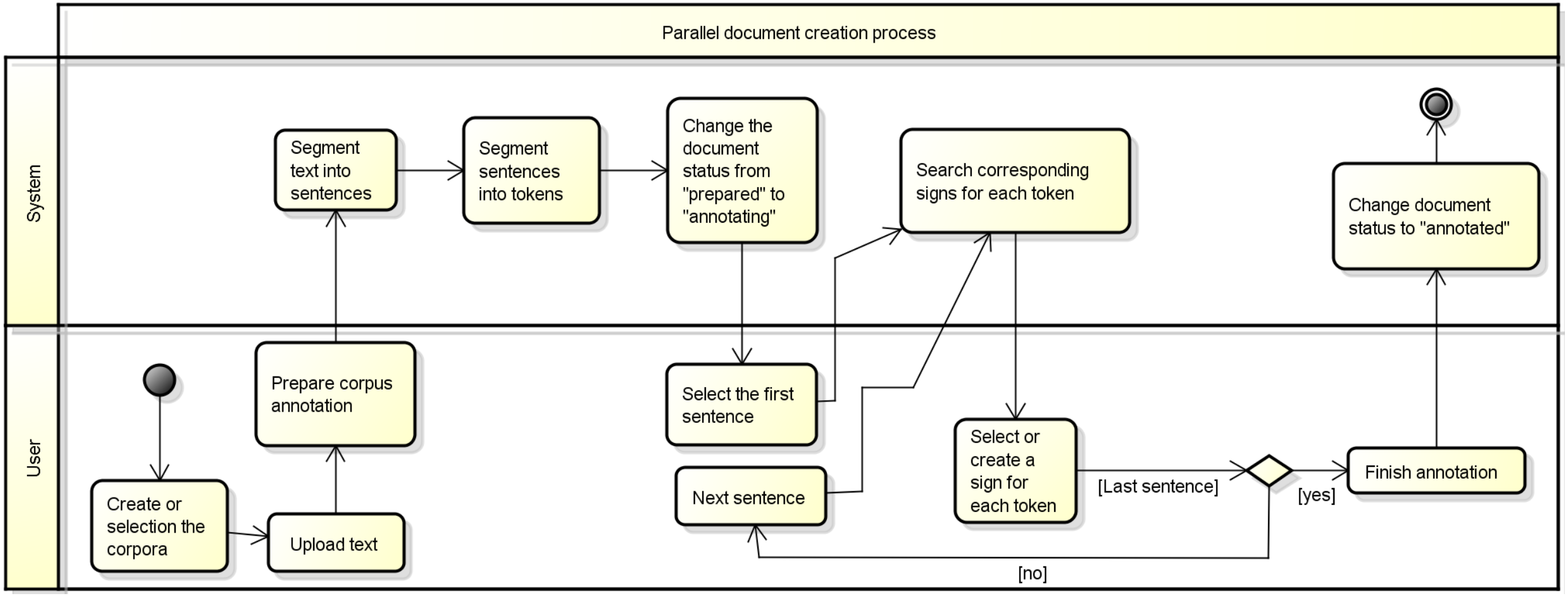
Design and Implementation:

- Java Web platform.
- EJB Application (Enterprise JavaBeans).
- JSF framework (Java Server Faces).
- MVC architecture (Model-View-Controller).

Diagram Class



Process creation of parallel corpus



SignCorpus annotator Home Corpora Dictionaries Help alex@porthal.com.br

Dictionary

New dictionary

+ Import dictionary

Dictionary of import in the format ".json".

Search:

(Pag. 1/1 - 2 records) 1 50

#	Language name	Language code	Number of signs	Action
1	Libras | ptBR | bzs	bzs	3735	
2	teste	teste1	0	

(Pag. 1/1 - 2 records) 1 50

SignCorpus annotator Home Corpora Dictionaries Help alex@porthal.com.br

Dictionary entries

Dictionary

Libras | ptBR | bzs

Search:

(Pag. 1/374 - 3735 records) 1 2 3 4 5 6 7 8 9 10 10

#	Word	Sign	FSW	Action
1	,		S38700463x496	
2	.		S38800464x496	
3	*a esceção de		AS10050S15a48S22a04S20e00M516x537S15a48485x462S10050501x470S20e00503x525S22a04503x504	
4	*alfabeto		AS14220S26904M520x521S14220481x491S26904504x479	
5	*alfabeto		AS14027S22a00M512x531S14027488x500S22a00490x468	
6	*alfabeto		AS14720M507x511S14720493x489	
7	*alfabeto		AS10020S2450aM525x515S10020475x485S2450a494x484	

SignCorpus annotator Home Corpora Dictionaries Help alex@porthal.com.br

Corpora

New Corpora

Search:

(Pag. 1/1 - 2 records) 1 50

#	Language name	Action
1	Libras | ptBR | bzs	
2	ASL | en_us	

(Pag. 1/1 - 2 records) 1 50

SignCorpus annotator Home Corpora Dictionaries Help alex@porthal.com.br

Sentence Annotate

Title

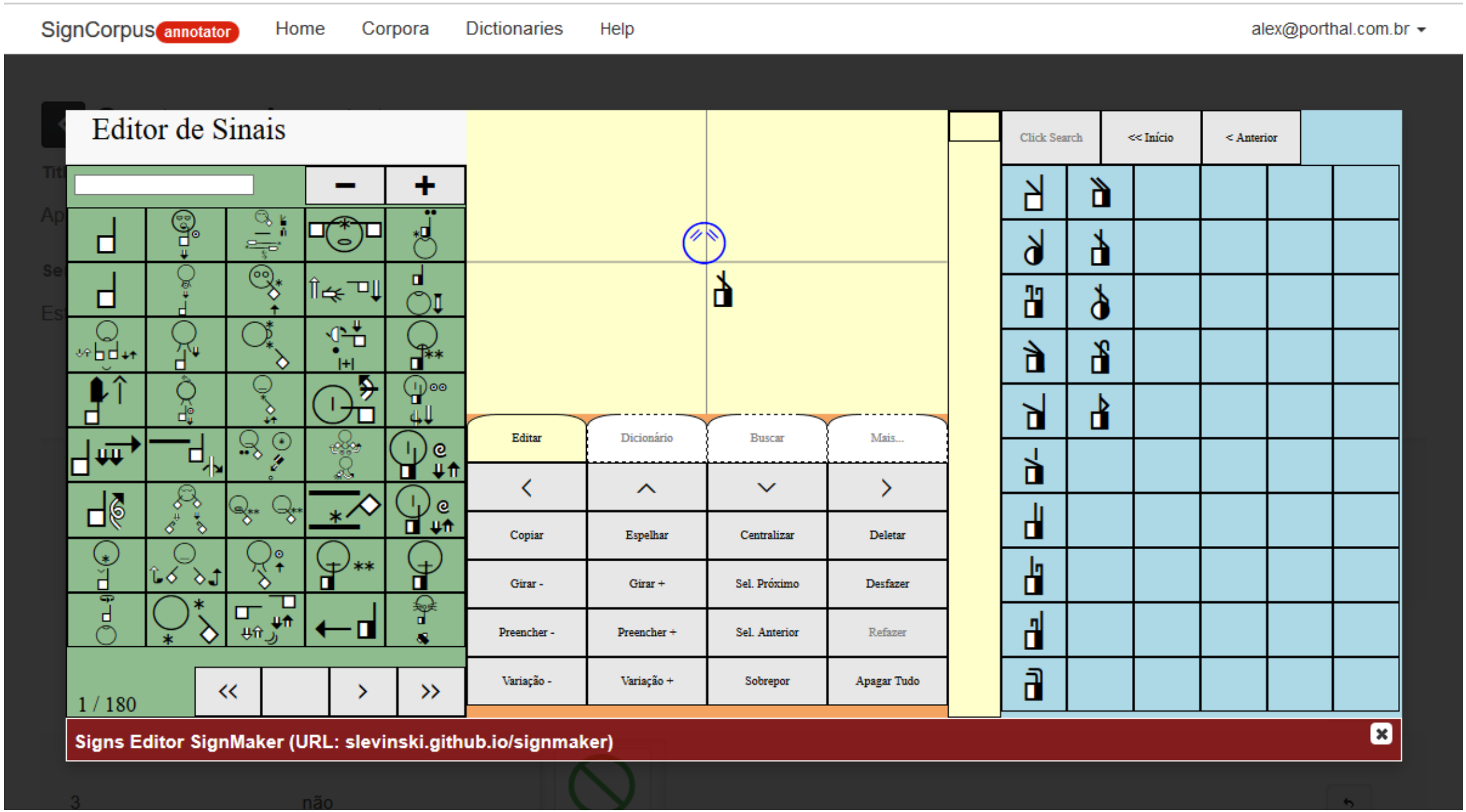
Aprendizado de máquina

Sentence

Esta página ou secção não cita fontes confiáveis e independentes, o que compromete sua credibilidade (desde junho de 2010).

(Pag. 1/1 - 21 records) 1 50

#	Operation	Word	Sign	Operation
1		Esta	Esta	
2		página ou secção	página ou secção	
3		não	não, nao, não, não, não, não	


SignCorpus annotator
Home
Corpora
Dictionaries
Help
alex@porthal.com.br
Editor de Sinais
1 / 180
Editar
Dicionário
Buscar
Mais...
Copiar
Espelhar
Centralizar
Deletar
Girar -
Girar +
Sel. Próximo
Desfazer
Preencher -
Preencher +
Sel. Anterior
Refazer
Variação -
Variação +
Sobrepor
Apagar Tudo
Click Search
<< Início
< Anterior
Signs Editor SignMaker (URL: slevinski.github.io/signmaker)

Schedule

First, a bit of history and context

Introduction

Theoretical foundation

Related Work

SignCorpus Annotator

Final Remarks and Future Work

Final Remarks and Future Work

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.

Open source: *https://bitbucket.org/unipampa/signcorpus*

Next step is to improve the searching and ranking of candidate signs by considering word inflections and by building language models for sign sentences.

Thank You!

Questions!?

;)