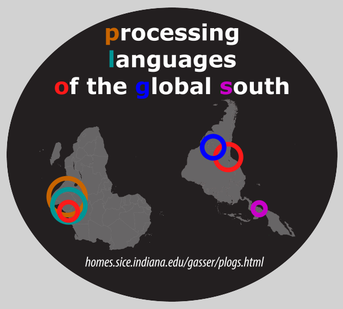
MAINUMBY: Computer-Assisted Spanish-to-Guarani Translation



Michael Gasser (*gasser@indiana.edu*)

Indiana University, School of Informatics, Computing, & Engineering (Emeritus)

MOTIVATION AND GOALS

Low-resource languages (LRLs) suffer from a lack of written material. Compare Finnish (472,303 articles) and Guarani (3,722 articles) Wikipedias.

Communities of translators exist for many LRLs, but **computer-assisted translation** (CAT) normally requires **bilingual corpora** (translation memories), which don't exist for LRLs, so translators can't benefit from technology.

Unlike machine translation (MT), CAT offers a translator options, desirable because of pervasive **ambiguity** and the lack of corpora for disambiguation.

Project goals:

- Framework for CAT systems translating from high-resource source languages (SLs) to low-resource target languages (TLs) using minimal resources: **Minimal Dependency Translation (MDT)**
- Implementation of MDT for translation between Paraguay's official languages, Spanish and Guarani

SYSTEM COMPONENTS

- **Bilingual lexicon** (LEX) of lexemes and phrases, with **cross-linguistic agreement constraints** (CLA)
- SL **morphological analyzer** (MA), TL **morphological generator** (MG)
- SL **morphosyntactic transformation rules** (MTRs) to make SL morphology resemble TL morphology
- Bilingual **morphosyntactic combination rules** (MCRs) to join lexicalized segments
- Minimal SL, TL, and bilingual **corpora** for limited disambiguation

PROCESSING IN MDT

- Source sentence (SS) is tokenized, POS-tagged, and morphologically analyzed.

  `"si no llueve"⇒si no llover[tns=prs,sbj=3s]`

- SS is matched against MTRs.

  `si no llover[…] ⇒ si llover[…,+neg]`

- SS is matched against LEX, resulting in a sequence of lexicalized segments, with CLAs to be satisfied. Multiple translations are ordered by frequency.

  `si llover[…] ⇒ SI LLOVER[…]~KY[] {tns=tns,neg=neg,sbj=sbj}`

- Sequences of bilingual segments are iteratively matched against MCRs.

  `SI LLOVER[…]~KY[] ⇒ LLOVER[…]~KY[pp=rõ]`

- CLAs are realized.

  `KY[pp=rõ] ⇒ KY[tns=prs,sbj=3s,+neg,pp=rõ]`

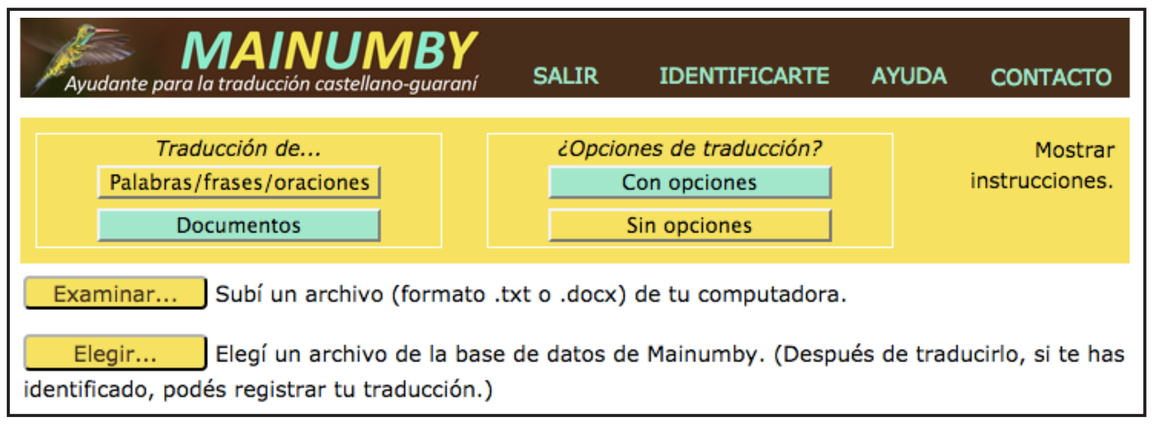
- TL wordforms are produced using MG. Multiple outputs are ordered by frequency.

  `KY[tns=prs,sbj=3s,+neg,pp=rõ] ⇒ ndokýirõ`

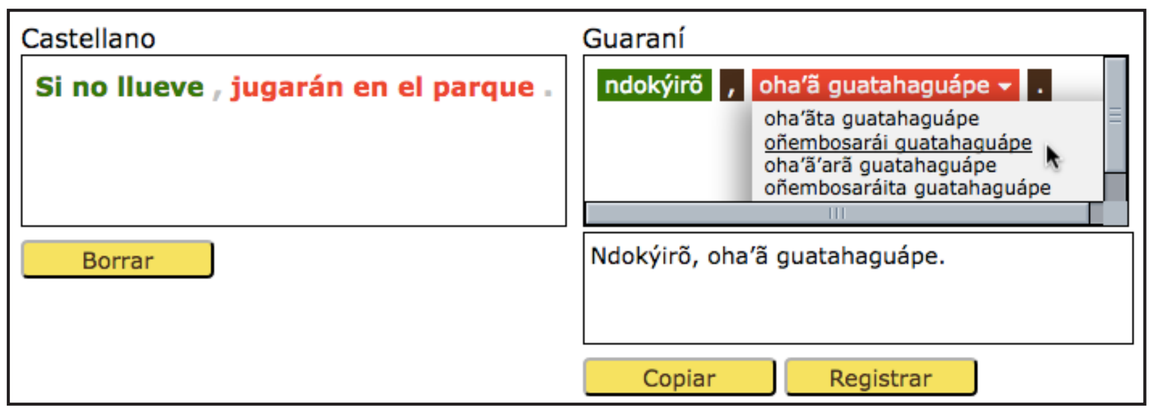
WEB APPLICATION

MAINUMBY (Guarani for 'hummingbird') is an implementation in Python of MDT for Spanish-to-Guarani translation (`github.com/hltdi/mainumby`). Translators interact with MAINUMBY through a web application (`plogs.soic.indiana.edu/mainumby/`).

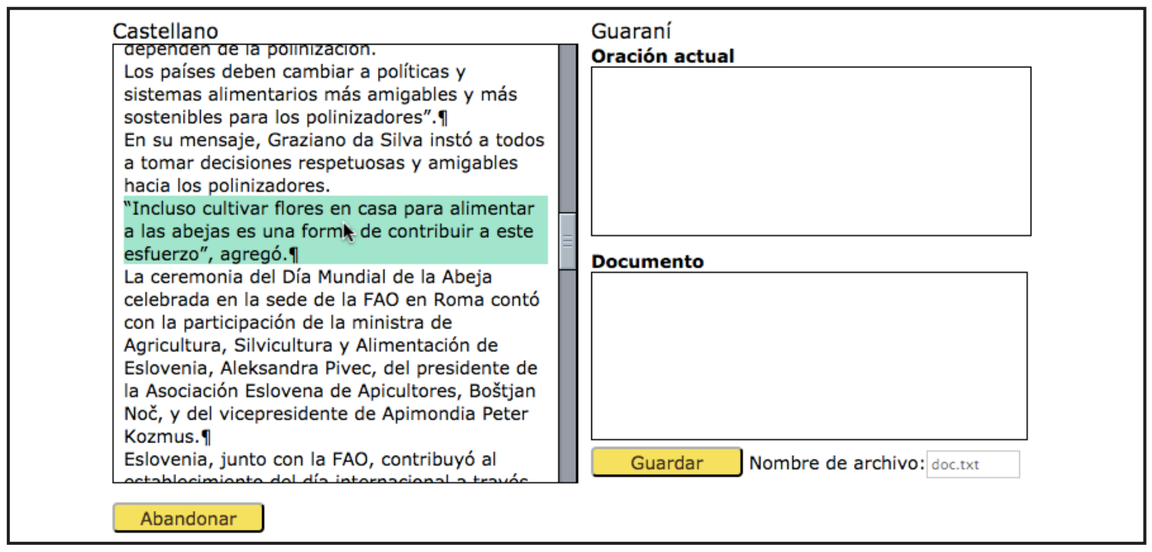
The web app features both sentence-level and document-level interfaces and translation offering options (as in CAT systems) or not (as in MT systems).



The segments resulting from MAINUMBY's translation are color-coded to represent cross-lingual correspondences. Users can re-arrange segments in the Guarani translation and can edit the translations they have selected before copying or recording them.



As in conventional CAT systems, the user may select a sentence within a document to translate.



STATUS AND ONGOING WORK

MAINUMBY includes 9,364 LEX entries, 89 MTRs, and 98 MCRs. It processes sentences at a rate of 23ms/word, leaving 28% of Spanish words untranslated.

User testing of MAINUMBY by bilinguals from the **Fundación Yvy Marãe'ỹ** in Paraguay is underway.

Learning from user feedback is being implemented.

MAINUMBY will be used as a baseline MT system for developing Spanish-Guarani NMT.