# FastSpell: the LangId Magic Spell

Marta Bañón (Prompsit Language Engineering)
mbanon@prompsit.com
LREC-COLING Torino (Italy), 20-25 May 2024

prompsit

# Table of contents

# 01

# Introduction

# 01. Introduction (I)

- Language identification: essential component.
- Three factors:
  - **Number of languages covered**
  - Accuracy
  - Speed

# 01. Introduction (II)

- What is FastSpell
  - Language identifier
  - Reviews and complements prior langid
  - Special focus: similar languages and language varieties
  - FastSpell = **fast**Text + Hun**spell**
- What does FastSpell
  - Focus on the **targeted language**
  - 1. Ask fastText
  - 2. Apply Hunspell
- Where is FastSpell
  - Bitextor & Monotextor pipelines
  - ParaCrawl, MaCoCu, HPLT

# 02 Why FastSpell?

# 02. Why FastSpell?

- ParaCrawl series
  - Deriving parallel data from web-crawled content
  - Language identification is mandatory
  - Web-crawled content = very noisy
- CLD2 issues:
  - Closely related languages
  - Uppercased text
  - Multi-script languages

# 03

# Benchmarking language identifiers

# 03. Benchmarking language identifiers (I)

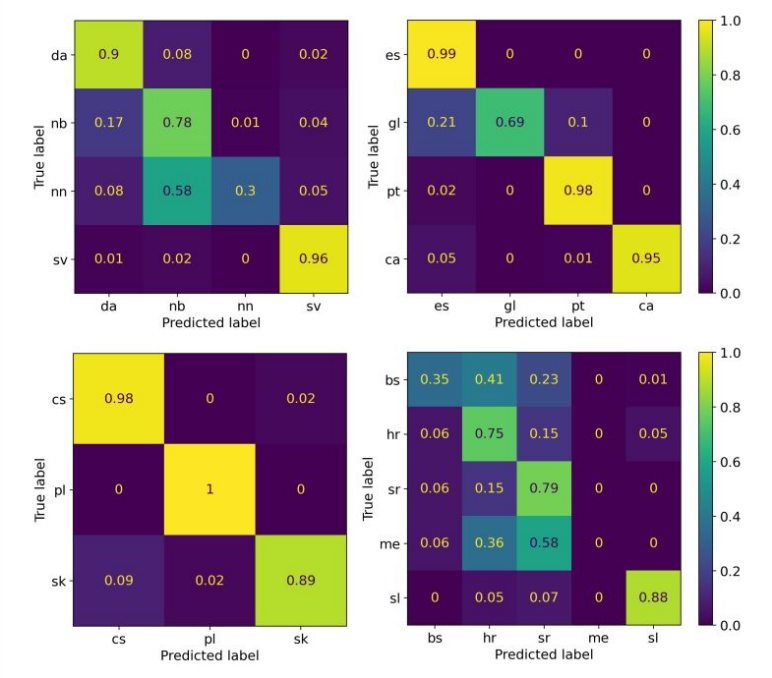
- Benchmark:
  - Performance (runtime)
  - Accuracy (F-score)
- Data:
  - ParaCrawl human annotation + SETIMES + Montenegrin gov.
  - https://github.com/mbanon/benchmarks
- Tools:
  - pyCLD2, pyCLD3, langid, FastLang, langdetect, NLTK, GuessLanguage, fastText, HeLI-OTS
  - pyCLD2+Hunspell, pyCLD3+Hunspell, fastText+Hunspell

# 03. Benchmarking language identifiers (II)

| Lang | F1 scores | | | | |
|---|---|---|---|---|---|
| | pyCLD2 | pyCLD3 | fastText 176 | HeLI | FastSpell |
| es | 0,890 | 0,932 | 0,929 | **0,957** | 0,954 |
| gl | 0,866 | 0,903 | 0,808 | **0,939** | 0,800 |
| ca | 0,904 | 0,935 | 0,951 | **0,961** | 0,935 |
| da | 0,914 | 0,868 | 0,824 | **0,930** | 0,799 |
| nb | 0,713 | - | - | **0,724** | 0,675 |
| nn | 0,781 | - | 0,433 | 0,787 | **0,810** |
| bg | 0,939 | 0,959 | 0,979 | 0,972 | **0,990** |
| cs | **0,965** | 0,908 | 0,918 | 0,941 | 0,962 |
| el | 0,976 | 0,964 | **1,000** | 0,991 | **1,000** |
| mk | 0,883 | 0,981 | **0,993** | 0,982 | 0,985 |
| ro | 0,960 | 0,947 | 0,970 | **0,988** | 0,975 |
| sk | **0,966** | 0,934 | 0,917 | 0,952 | 0,937 |
| sl | 0,925 | 0,887 | 0,847 | **0,942** | 0,880 |
| sq | 0,981 | 0,986 | **0,993** | 0,992 | 0,990 |
| mt | 0,984 | 0,976 | 0,896 | **0,996** | 0,914 |
| tr | 0,978 | 0,985 | **0,990** | 0,983 | 0,988 |
| bs | **0,443** | 0,336 | 0,416 | - | 0,370 |
| me | - | - | - | - | **0,458** |
| hr | **0,641** | 0,423 | 0,557 | - | 0,541 |
| sr | 0,529 | 0,327 | **0,565** | - | 0,493 |
| hbs | 0,940 | 0,941 | 0,922 | 0,982 | **0,983** |
| Avg. Runtime | **0,011** | 0,097 | 0,019 | 2,688 | 1,076 |

- Results: https://tinyurl.com/2u48kycz
- pyCLD2:
  - Best candidate (runtime)
  - Already used in the pipeline
- HeliOTS:
  - Best candidate (F-score)
  - 100x slower than fastText
- **fastText**
  - 2nd best F-score
  - 2nd best runtime

# 03. Benchmarking language identifiers (III)




- fastText:
  - Issues with similar languages
    - nn - nb
    - gl - es
  - Inaccurate predictions
    - sk
  - Unsupported languages
    - me

# 04
# The FastSpell spell

# 04. The FastSpell spell (I)

- FastSpell = **fast**Text + Hun**spell**
- fastText
  - Library for fast text representation and classification
  - Free/open source
  - Developed by Meta
  - Model: lid.176.bin (176 langs)
- Hunspell
  - Spell checker and morphological analyser
  - Free/open source
  - Uses affix files and dictionaries

# 04. The FastSpell spell (II)

- Focus on the **targeted language**
- Predict language with **fastText**
- If prediction is the targeted language or a similar one:
  - Refine prediction with **Hunspell**
    - Targeted language
    - Predicted language
    - Other similar languages
  - Select the candidate with less spelling errors
- Final prediction depends on the **mode**
  - **Aggressive**: More prone to choose the targeted language
  - **Conservative**: More hesitant to give predictions, can tag as **unknown**

# 05 Using FastSpell

# 05. Using FastSpell (I)

- Parameters:
  - Text to be identified
  - Targeted language
  - Mode

**Python package**

```
>>> fs_gl = FastSpell("gl", mode="aggr")
>>> fs_pt = FastSpell("pt", mode="aggr")
>>> fs_de = FastSpell("de", mode="aggr")
>>> fs_gl.getlang("Nunca choveu que non escampara")
'gl'
>>> fs_pt.getlang("Nunca choveu que non escampara")
'pt'
>>> fs_de.getlang("Nunca choveu que non escampara")
'pt'
```

**CLI tool**

```
#echo "Nunca choveu que non escampara" | fastspell gl --aggr
Nunca choveu que non escampara  gl
# echo "Nunca choveu que non escampara"| fastspell pt --aggr
Nunca choveu que non escampara  pt
# echo "Nunca choveu que non escampara"| fastspell de --aggr
Nunca choveu que non escampara  pt
```

# 05. Using FastSpell (II)

## Out-of-the-box

- Preconfigured for several languages
- Includes resources
- Covers ParaCrawl/MaCoCu/HPLT languages

## Custom configured

- Achieved by modifying config files
- similar.yaml
  - targeted languages and their similar languages
- Hunspell.yaml
  - Hunspell dictionaries to be used

06

# Future work

# 06. Future work

## FastSpell

- Tokenization/stemming
- Non-targeted identification
- Different thresholds

## fastText

- lid.176.bin replacements
- Faster implementations

## Hunspell

- Build more dictionaries
- Write multi-lang engine

# Acknowledgement

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant no 10052546].


UK Research and Innovation

# FastSpell: the langid magic spell

github.com/mbanon/fastspell
pypi.org/project/fastspell

Presentation template created by Slidesgo

prompsit

FastSpell logo created by DailyPM from Noun Project

# FastSpell: the langid magic spell

Marta Bañón (Prompsit Language Engineering)
mbanon@prompsit.com
LREC-COLING Torino (Italy), 20-25 May 2024