How Does Language Influence Documentation Workflow?

Unsupervised Word Discovery Using Translations in Multiple Languages

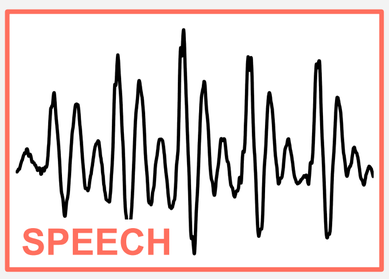
Marcely Zanon Boito, Aline Villavicencio, Laurent Besacier
contact: marcely-zanon-boito@univ-grenoble-alpes.fr

1. Context: Computational Language Documentation (CLD)

- More than 50% of the currently spoken languages will **no longer exist** by the end of the century [1]
- Manually documenting all these languages is **infeasible**

CLD GOAL: to automatically retrieve information about language structures to speed up language documentation

Many of these **endangered languages** are unwritten! In this scenario, *translations* replace *transcriptions* of the recordings [2]

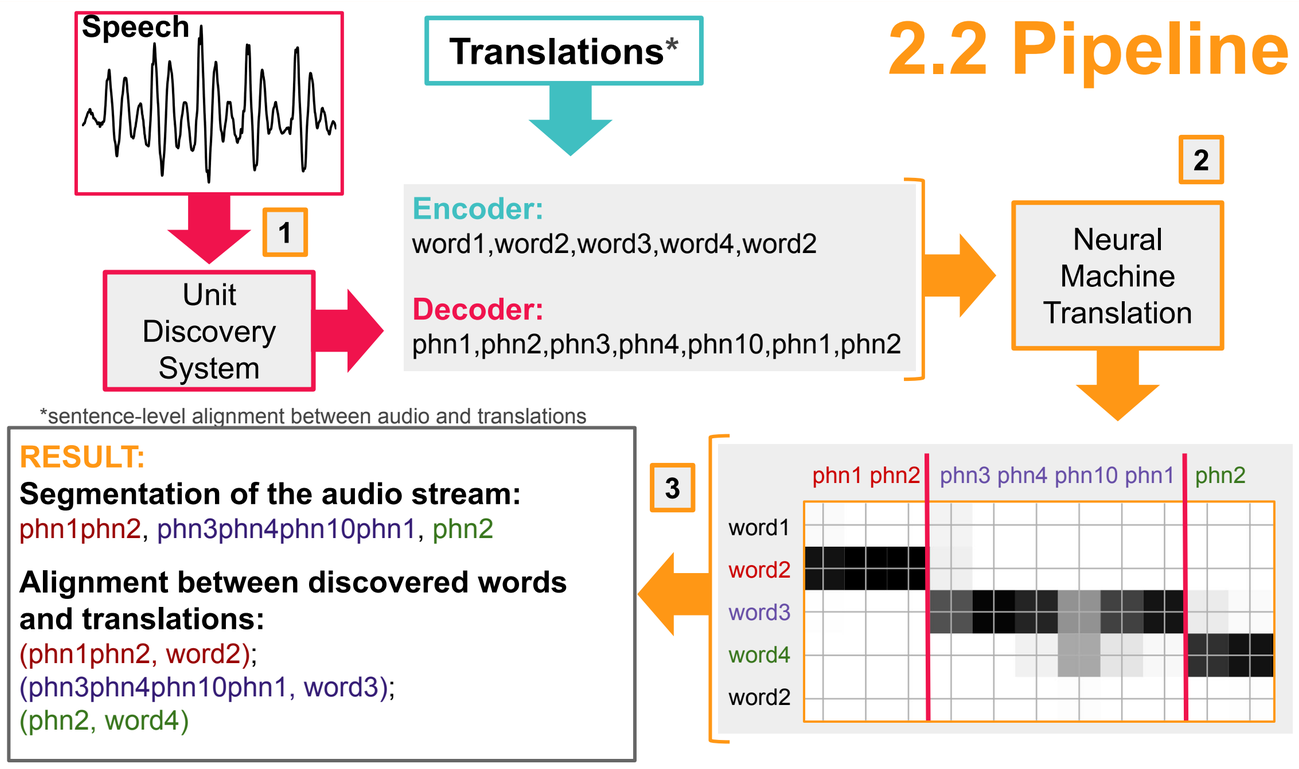


2. Method: Unsupervised Word Segmentation/Discovery (UWS)

2.1 Method [3]

1. **[speech2phone]** Speech in the endangered language is encoded as a phone sequence by an automatic unit discovery system; **In this paper we use a topline: phones transcribed by a linguist.**
2. **[alignment]** Translations and unsegmented phone sequences (sentence-level aligned) are fed into a *Neural Machine Translation (NMT) System*;
3. **[segmentation]** The **soft-alignment** learned by the NMT system is used for segmenting the phones sequence into word-like units. Alignment between discovered and translation words is also retrieved.

2.2 Pipeline



3. How Does Language Influence Documentation Workflow?

3.1 Motivation

Different language pairs might capture different optimal source-to-target correspondences, thus:

- Segmentation performance might depend on the **aligned information (language used for documentation)**
- Combining information from different bilingual models might enrich our segmentation **(multilingual setup)**

3.3 Results and Lessons Learned

We train five bilingual models for **segmenting Mboshi phones**, each one using a different source (documentation) language. We verify a **small performance difference** comparing these models, with less similar languages scoring worse (Table below). Combining the models for generating a multilingual-rooted segmentation marginally increased boundary results (74.9%);

	Types	Boundary
FR	27.6	73.4
EN	27.7	73.1
PT	27.6	72.8
ES	26.6	72.6
DE	24.0	71.0

Table: F-score results for UWS using bilingual models.

- Similar languages score better, but results may be limited by the automatic generated translations;
- EN results might be explained by statistical features; (see Table at 3.2)
- Results hint at chosen documentation language impacting performance for bilingual UWS.

3.2 Languages

We extend the existing bilingual **Mboshi (MB)** and **French (FR)** documentation corpus [5] by translating the well-resourced French into four other languages using the DeepL translation platform*: **English (EN)**, **Spanish (ES)**, **German (DE)** and **Portuguese (PT)**. Mboshi (Bantu C25) is an unwritten language spoken in Congo-Brazzaville and documented by the BULB project. [2,4]

*https://www.deepl.com/translator

	MB	FR	EN	ES	DE	PT
# Types	6,633	5,178	4,392	5,473	5,641	5,465
# Tokens	30,556	42,715	37,379	37,428	37,515	37,095
Avg Token Length	4.18	4.41	4.19	4.36	4.91	4.40
Avg Tokens/ Sentence	5.96	8.33	7.29	7.30	7.31	7.23

Table: Statistics for the multilingual corpus used in this investigation.

References

[1] Austin, and Sallabank, (2011). ***The Cambridge handbook of endangered languages.*** Cambridge University Press.
[2] Adda et al. (2016). ***Breaking the unwritten language barrier: The BULB project.*** SLTU 2016.
[3] Boito et al. (2019) ***Empirical Evaluation of sequence-to-sequence models for word discovery in low-resource.*** Proceedings Interspeech 2019.
[4] Stüker et al. (2016) ***Innovative technologies for under-resourced language documentation: The Bulb project.*** CCURL 2016.
[5] Godard et al. (2018) ***A very low resource language speech corpus for computational language documentation experiments.*** LREC 2018