SADiLaR
South African Centre for Digital Language Resources

Corpus-based dictionaries for low-resource languages

Mrs Mmasibidi Setaka-Bapela
Prof Menno Van Zaanen

Funded by:

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Introduction

- Dictionaries are useful lexicographic resources
- Dictionaries
  - come in different forms, e.g., paper, electronic, pocket
  - are great language learning aids
  - help with language preservation and language documentation
  - are designed to assist users in language-related tasks
- However, not all languages have dictionaries
  - Out of +- 7000 languages of the world, only a fraction have lexicographic resources
  - A large majority of languages have limited or no lexicographic resources

High-Resource Languages (HRLs)

- HRLs have many computational resources, e.g., English
- HRLs receive more (research) attention and a sense of importance
- HRLs have many low-level language resources, such as corpora, for the building and development of (high-level) resources
- HRLs typically have dictionaries in different formats that are continuously updated

Low-Resource Languages (LRLs)

- LRLs have no or very limited computational resources
- LRLs do not necessarily have few speakers
- LRLs are not second (or third) rated languages (!)
- If LRLs have computational resources, such as corpora, they are generally
  - of limited size
  - of limited variability (e.g., genres)

Developing dictionaries for LRLs?

- Most low-resource languages do not have (enough) corpora

> "Valuable and even sufficient data for the compilation of a specific dictionary can be extracted from a relatively small corpus of approximately one million words."
>
> (Prinsloo 2015)

- How suitable are corpus-based methodologies for the compilation of dictionaries of low-resource languages?

Current trends in lexicography

- Corpus-based approaches to dictionary creation are highly suitable for HRLs
  - These languages have large amounts of digital language resources
- Lexicographers require access to empirical data for accurate information
  - Corpora have become the preferred source of evidence
- Dictionaries of the future will rely heavily on electronic corpus-based studies
  - E.g., Oxford Bilingual School Dictionary: IsiZulu \& English (2014) is corpus-based

Challenges for LRLs

"It seems likely that by the middle of this century, if not before, all dictionaries will be in electronic form."

(Oxford Dictionaries 2024)

- Quality and availability of corpora
  - Without corpora, the lexicographer relies on intuition
- Non-language resource constraints: time, skills, personnel, etc.
- Many people speaking LRLs do not necessarily have access to online material
  - Internet access, mobile phones, computers, etc.

South African situation

- Many South African languages have limited (digital) language resources
  - Small corpora
  - Unbalanced corpora
  - Non-representative corpora
- Limited corpus availability limits research for these languages (not just for dictionaries)
- (For non-official languages, the situation is typically even more dire)
- Potential solutions
  - SWiP project (aims at getting more language data in Wikipedia for South African languages)
  - Digitization projects
  - ...

Additional challenges (specific to LRLs)

- How to identify neologisms (newly coined words or expressions)?
  - If only limited or non-recent corpora are available
- At what stage are words (e.g., neologisms) included in dictionaries for LRLs?
  - How and when to update LRL dictionaries?
  - Are Coronavirus wordlists (developed and circulated in South Africa) now included in dictionaries?
- What is a dictionary (versus word list)?
  - Many languages have (multilingual) wordlists; are these dictionaries?

"In the African context “dictionary” is a relative term,
since mere word lists are often regarded as dictionaries."
(Prinsloo 2015)

Preliminary results

- Small or limited corpora mean insufficient examples of language use are available
- Non-corpus-based approach may lead to
  - limited vocabulary in dictionary
  - words that are currently in use missing in dictionary
  - presence of obsolete words (words no longer in use)
- Language properties may influence required size of the corpus, e.g., morphological complexity

Focus points

- Compare traditional versus corpus-based approaches to dictionary creation
  - To understand the impact on dictionary development for LRLs
- Provide overview of available tools (computational or non-computational)
- Investigate
  - corpus properties that allow for corpus-based dictionary development
  - language properties that allow for corpus-based dictionary development
- Provide measure of suitability of corpus-based approach for the current state of LRLs

Methodology

- Literature review
- Interviews, focus groups, surveys looking at:
  - dictionary processes and tools
    - corpus-based and non-corpus based approaches
    - computational and non-computational
  - non-tool resources
  - language properties
- Create a baseline for the quantity and availability of resources for these languages

Ke a leboha

Mmasibidi Setaka
Mmasibidi.Setaka@nwu.ac.za

SADiLaR
South African Centre for Digital Language Resources