Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL), pages 9–14
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Usability and Accessibility of Bantu Language Dictionaries in the Digital Age: Mobile Access in an Open Environment

Thomas Eckart, Sonja Bosch, Uwe Quasthoff, Erik Körner, Dirk Goldhahn, Simon Kaleschke
Natural Language Processing Group, University of Leipzig, Germany
Department of African Languages, University of South Africa, Pretoria, South Africa
{teckart, quasthoff, koerner, dgoldhahn, skaleschke}@informatik.uni-leipzig.de
boschse@unisa.ac.za

Abstract

This contribution describes a free and open mobile dictionary app based on open dictionary data. A specific focus is on usability and user-adequate presentation of data. This includes, in addition to the alphabetical lemma ordering, other vocabulary selection, grouping, and access criteria. Beyond search functionality for stems or roots – required due to the morphological complexity of Bantu languages – grouping of lemmas by subject area of varying difficulty allows customization. A dictionary profile defines available presentation options of the dictionary data in the app and can be specified according to the needs of the respective user group. Word embeddings and similar approaches are used to link to semantically similar or related words. The underlying data structure is open for monolingual, bilingual or multilingual dictionaries and also supports the connection to complex external resources like Wordnets. The application in its current state focuses on Xhosa and Zulu dictionary data but more resources will be integrated soon.

Keywords: dictionary data, mobile application, usability

1. Introduction

Lexical data sets are an indispensable resource for a variety of user groups, ranging from school children to professional text creators. However, the traditional ways of presenting and distributing this valuable knowledge by means of printed books does not reach all potential users anymore. New ways of data access and participation have to be identified and implemented as part of their further development. Even though many relevant resources are already available via Web pages, recent trends to extended use of dedicated mobile applications (apps) especially by a younger audience are only considered to a small extent and have led to - if any - a variety of incompatible, proprietary and therefore - after some time - abandoned applications with unmaintained data stocks.

The mobile cellular community in Africa is a fast growing one. In the case of South Africa, it was reported by Statistics South Africa[1] that the proportion of households owning mobile phones significantly increased from 31.9% in 2001 to 88.9% in 2011, while a community survey in 2016 (Statistics South Africa, 2016) indicated a further increase to 93.8% of households. Mobile phones resorted under the category “household goods”, and interestingly enough, achieved the highest percentage after electric stoves, TVs, and fridges. Mobile versions of Bantu language dictionaries could therefore facilitate accessibility to a large percentage of the population in contrast to traditional dictionaries which are expensive, often out of print and even outdated. Such electronic dictionaries also save users time compared to paper dictionaries. Moreover, they “save working-memory for comprehension processing rather than being disrupted by taking much time finding words in traditional dictionaries” (Deng and Trainin, 2015:58).

Taking this general environment into consideration, this contribution focuses on an Android dictionary application designed as an open source project to enhance the visibility of available resources and as an attempt to reach and activate new user groups. It will be shown how available resources - in part compiled or prepared by the authors themselves - can be made accessible and how openness can help to achieve similar results for other resources as well. Based on the analysis of existing mobile applications and their shortcomings, some approaches to improve the presentation and accessibility of data on a limited screen will be depicted with a focus on (semi-)automatic approaches for less-resourced languages.

1 http://www.statssa.gov.za

2. Openness as Prerequisite for Collaboration and Participation

The FAIR data principles (findability, accessibility, interoperability, and reusability; see Wilkinson et al., 2016) have a growing influence on the everyday work of researchers and scientists. However, this - in general accepted - focus on a minimal set of requirements for allowing modern and open research is still not implemented in all areas. The consequences are serious and problematic especially for disciplines where the availability of reliable resources itself is problematic. Among others, this is specifically the case for many African indigenous languages of which most can be considered as resource scarce.

To achieve an open environment where interested researchers and users can collaborate and develop resources continuously, the required level of “openness” does not only include the ability to find, access, interoperate, and reuse data. In the context of this contribution, the focus lies on a more complete scenario when providing access to lexical resources for Bantu languages. The following views on openness are of particular relevance here:

- Open data: The availability of data for research, aggregation, and for re-use in other contexts is an obvious prerequisite for an active community and continuous development of the language

resources landscape. In this contribution, a Xhosa dictionary dataset that was previously made available by the authors under an open license (Bosch et al., 2018), is used. However, this only serves as a concrete example; a limitation to this specific dataset or Bantu language is not intended. The openness of data and compliance with general standards of their formal representation allow the integration of other resources as well, as already tested using resources from the Comparative Bantu Online Dictionary project (CBOLD[2]).

- Open application: Besides the focus on data respecting the FAIR principles, the reusability of applications is another important aspect. Open or free software[3] allows the reuse of applications for new purposes or data sets and their collaborative development and improvement. Therefore, the application presented here is made freely available[4] under an open source licence and can be reused by other interested parties.

- User-friendly application: The open availability of data via open user interfaces is only one prerequisite to attract users and potential collaborators. The user experience provided by an application and its appropriateness for relevant user tasks is another important precondition. Unfortunately, most of the current applications - including commercial apps - are only trying to transfer established paradigms of structuring and accessing dictionary data to the digital age. The following sections will focus on new approaches that are still feasible for the problematic field of less-resourced languages.

- User-friendly data import: To simplify the re-use of the application, the effort that is necessary to import other data sets should be kept as low as possible. This can be achieved in different ways: by relying on established standard formats and/or by providing a simple mechanism to feed data into the application that is applicable even for inexperienced users. The authors have decided to select a dual approach in which lexical data is provided in form of column separated value (CSV) files which can be created, maintained, and edited using standard office software (like LibreOffice or Microsoft Excel). For more elaborate and established formats, transformation procedures are provided. This currently includes transformation scripts for data structured according to the Bantu Language Model (BLM) which is based on the MMoOn ontology (Klimek, 2017). The support of additional formats is planned for the future.

- External open resources: No application can provide all available information for a language or incorporate all established and often very extensive external resources. However, direct links are a helpful feature and make use of the distributed landscape of language resources. In this context, referencing data of the African Wordnet (AfWN) (Bosch and Griesel, 2017) and the dynamic incorporation of extracted full-text material as usage samples (Goldhahn et al., 2019) via RESTful Web services (Büchler et al., 2017) is considered to be of high relevance.

Current endeavours towards integrated and open research infrastructures like the South African SADiLaR[5], the European CLARIN/CLARIAH (Hinrichs & Krauwer, 2014) and more can be seen as the natural context for all of these developments.

3. User Groups and Profiles

This general complexity of data access when based on a relatively simple data format should not restrict usability requirements. There is a variety of potential user groups such as language learners of different ages (pupils of different ages, adults), different skill levels (beginners, L1 and L2 learners, professionals), different tasks to accomplish (text reception vs. text creation), and different types of dictionaries (monolingual, bilingual or multilingual with different amount of details).
A single dictionary may address a single user group or multiple user groups. The combination of targeted user group and available data in the dictionary determine dictionary details presented to the user. For a given dictionary, the presentation for different user groups is defined by the dictionary data provider within the dictionary, ideally together with the dictionary author(s). The result is either a single interface option for a given dictionary or a selection of two or three different interfaces (for instance, for beginners or professionals), where the user can select the appropriate option. The interface definition applies both to macro- and microstructure. The macrostructure should be accessible to pre-select the lemmas shown to the user. In a usual dictionary, all lemmas are presented in alphabetical order. On this level, we have the option to restrict the set of lemmas (for instance, for beginners or to focus on specific subject areas) and to change their order.
For the microstructure, we can restrict the dictionary by ignoring some information which is assumed to be known to (or irrelevant for) the targeted dictionary user. This may include information in bilingual dictionaries which are in the user's mother tongue or information irrelevant for the specific task that the user tries to accomplish.

As a result, defined user groups have to be aligned to supported user profiles with direct consequences for the selection and presentation of lexical data. This alignment may be structured according to the following examples:

- For language learners, it is highly relevant to access words belonging to the same semantic field in a combined presentation. This may include vocabulary which is part of the same semantic field or - especially in the context of

2 http://www.cbold.ish-lyon.cnrs.fr

3 For the following definition of "free software": https://fsfe.org/about/basics/freesoftware.en.html

4Available at https://github.com/cheapmon/balalaika

5 https://www.sadilar.org

primary education - part of the same lesson. The selection of presented lemmas is also defined by users' abilities and might include the restriction to high frequent terms, basic vocabulary, or terms known from previous lessons. On the microstructure level, this might comprise a focus on translations and concrete usage examples, while reducing the amount of morphosyntactic information to a minimum.

- For professional writers a suitable user profile can be constructed accordingly. This might also include an exclusive focus on domain-specific vocabulary (omitting basic vocabulary completely), taxonomic information (like synonyms or antonyms), and references to external, additional sources for non-lexical information.

This focus on user profiles might be seen as an unnecessary restriction in comparison with the absolute flexibility of a user-driven configuration. However, the willingness of users to adapt an interface to their specific needs is often low, which is in clear contrast to the technical costs of providing this flexibility in an application. The reduction of options to a reasonable subset is seen by the authors as a viable compromise.

3.1 Approaches for Lemma Selection

Typically, on a smartphone display, a maximum of ten lemmas can be presented in addition to a selected dictionary entry. The selection of these lemmas is crucial for easy dictionary use. The standard solution is the selection of the alphabetically neighboring words in the lemma list. In many cases, there are more attractive alternatives:

- Alphabetical subselection by frequency: In a large lemma list, many infrequent words are contained. Especially a language learner might be interested in medium or high frequency words only.
- Alphabetical subselection by the dictionary compilers: Words may be marked by difficulty (as beginners vocabulary, for instance), or subject area (medicine, for instance). Each subset can be selected, and all other words are ignored in the lemma list.
- Semantically similar or related words instead of alphabetic order: Semantically related words can either be provided by the dictionary (as by Wordnet, for instance) or generated automatically by word embeddings like Word2Vec or similar approaches. See the following section for more details.

4. Lemma Selection Approaches for Less-resourced Languages

The approaches identified for an improved access and presentation of lexical data would typically rely on extensive, mostly manually created resources. This includes vocabulary lists for specific domains (like vocabulary relevant for different school lessons or fields of work) or extensive taxonomic data. Unfortunately, for less-resourced languages those are not always available.

One of the positive developments is the recent effort on creating African Wordnets[6]. Linking up with a Wordnet provides additional suggestions such as synonyms or related concepts, definitions and usage examples in order to provide more learning opportunities. The African Wordnets project is currently under development for nine Bantu languages spoken in South Africa. Currently the prototypical African Wordnet (AfWN) contains open source data of varying sizes for the nine official African languages of South Africa. The AfWN is closely aligned with the English Princeton WordNet (PWN)[7] which forms the basic structure for continual and manual expansion of the AfWN (Bosch and Griesel, 2017). This so-called expand method offers an established structure for building a new resource and is therefore usually preferred for less-resourced languages (Ordan and Wintner, 2007:5). This method requires translation of the PWN into the target African language.

There is also a variety of statistics-based approaches to enhance dictionary usability for the purposes identified above. Most of these can be seen as semi-automatic procedures that are able to generate candidates but still require manual inspection and approval. Currently, the following approaches are evaluated with respect to the problem of data sparseness that applies to all less-resourced languages.

4.1 Differential Wordlist Analysis

The analysis and comparison of word lists (Kilgarriff, 2001) has proven to be useful for a variety of applications, including the corpus-based extraction of domain- or author-specific vocabulary (Goldhahn et al., 2015). This can be used for the purposes sketched in this contribution as well.

As a specific show case, vocabulary was identified that is suitable for primary school children. The used approach relies on the comparison of relative word frequencies in domain-specific texts compared with the frequency in a more general reference text corpus. As domain-specific material, texts of the *Nal'ibali*[8] project were used. Nal'ibali is a campaign to promote a reading culture in South Africa and provides multilingual stories in 11 languages. A word list was generated using the Zulu texts (around 34,000 tokens) and compared with a reference corpus of around 15 million tokens provided by the Leipzig Corpora Collection (Goldhahn et al., 2012) that aggregates text material using Web crawling for hundreds of languages.

6 https://africanwordnet.wordpress.com
7 https://wordnet.princeton.edu
8 https://nalibali.org

The resulting word form list contains both function words and everyday vocabulary which can be used for vocabulary selection. As concrete examples, the following inflected terms were extracted: kakhulu (*very much*), umuntu (*person*), kusho (*say/mean*), ukudla (*food*), umama (*mother*), ubaba (*father*), izilwane (*animals*), unogwaja (*rabbit*). Figure 1 and 2 compare the presentation for ilanga (*sun*, *daytime*) in an alphabetical order using a complete Zulu dictionary (thus including unrelated lemmata having the same prefix) with its presentation among a subset of vocabulary, extracted from the same source.

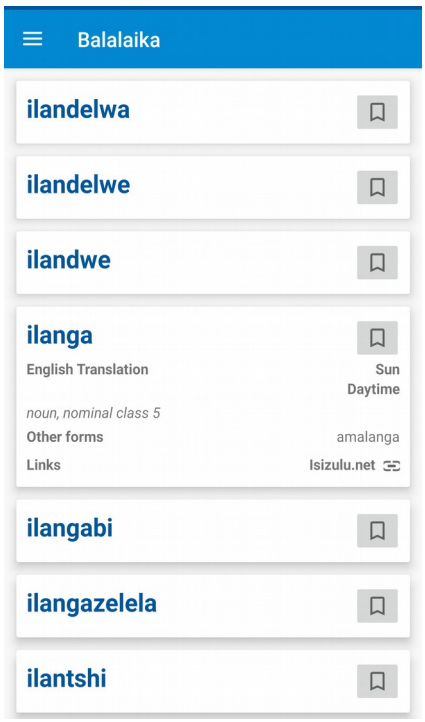

Figure 1: Zulu *ilanga* in an alphabetical lemma list

The sketched approach is of course not only usable for this specific kind of material, but can be applied to other genres as well. The basis in every case is a word list extracted from domain-specific texts. This can include school books, technical manuals, selections of Web pages or any other kind of text material.

4.2 Word Embeddings

The usage of word embeddings like Word2Vec (Mikolov et al., 2013) and Fasttext (Bojanowski et al., 2017) allows a variety of enhancements when using digital lexica. Their primary feature to compute semantic and/or syntactic similarity between two words can be used to provide different grouping options (clustering based on topic or similarity), suggestions of semantically related words and enables searching even with misspelled input words (Piktus et al., 2019). Word embeddings are comparable to word co-occurrences as they are both methods that exploit word contexts to "learn" the meaning of a word and related words. They are slightly more efficient to compute compared to traditional co-occurrences and can be stored more compactly which is helpful considering the limited amount of storage capacity on mobile devices.

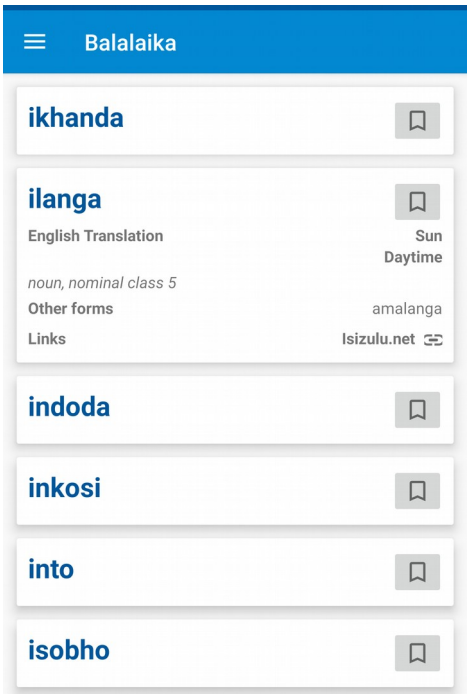

Figure 2: Zulu ilanga in a selection of domain-specific lemmata based on Nal'ibali texts (including *head*, *man*, *king*, *thing*, *soup*)

Different embedding techniques and models allow a choice between types of similarity, with Word2Vec focusing more on a kind of semantic similarity based on shared contexts, whereas Fasttext results are more morphologically similar as their calculation includes character n-grams, which is advantageous when working with unknown and infrequently occurring words. The choice of model can depend on language characteristics and user profiles. If possible, models should be trained on all available textual data for the given language. But even rather small text collections of about one million sentences allow for good results as shown below for Zulu. Training with even less data is possible but quality increases with more training examples.

A negative aspect for offline app usage is that the model size even for rather small corpora is between 100 MB and 1 GB and grows with the vocabulary size (related to the text corpora size). To minimize the initial app size, improvements such as pre-computing a fixed number of similar words for each vocabulary entry, on-demand downloading of (larger) models or word lists or even compressing embedding vectors (Joulin et al., 2016; Shu et al., 2017) can help mitigate this issue.

As concrete examples, the following two lexical items and their most similar forms in the dictionary according to (sub-) word similarity using Fasttext are provided. The

results are based on a 1.1 million sentences Zulu corpus[9]; English translations are provided in brackets.

- **ukulangazela** (*to long for*): nokulangazelela (*and longing*), nokukulangazelela (*longing for you*), ukulangazela (*longing*), unokulangazelela (*you can look forward*), kunokulangazelela (*more longing*), enokulangazelela (*longing*), Ukulangazelela (*Longing*), ukulangazelela (*longing*), yikulangazelela (*look forward to it*), Ukulangazelele (*You longed for it*), wokulangazela (*of longing*), njengokulangazela (*as longing*), okulangazelele (*that longed for*), kulangazela (*longing*), akulangazelelayo (*that/which/who long for it/you*), ukulangazelele (*you long for it*), ikulangazelela (*he/she/it/ longs for it*), ezokulangazelela (*that will long for you/it*), engakulangazelela (*that can long for it/you*)

The above examples represent inflection of the same basic verb stem by means of a variety of affixes. The meaning of the intransitive verb stem -langaza (*have a longing*) is extended by so-called verbal extensions -el- and -elel- to change the meaning to a transitive one, i.e. -langazela (*long for*). Various prefixes feature in these examples, ranging from the infinitive noun class prefix uku- in the word ukulangazela *(longing/to long for*) to subject and object agreement morphemes i- and -ku- in the word ikulangazelela (*he/she/it longs for it*) and possessive morphemes as in wokulangazela (*of longing*), to mention a few.

- **ibhayoloji** (*biology*): ibhayotheknoloji (*biotechnology*), ibhayografi (*biography*), iMayikhrobhayoloji (*Microbiology*), ezebhayoloji (*of biological*), yibhayotheknoloji (*it is biotechnology*), Ibhayotheknoloji (*Biotechnology*), bhayotheknoloji (*biotechnology*), ngeradiyoloji (*with radiology*), ifonoloji (*phonology*), ithayithili (*title*), nakwibhayoloji (*and in biology*), kwebhayoloji (*of biology*), nethayithili (*and a title*), kwemayikhrobhayoloji (*of microbiology*), zebhayoloji (*of biological*), ngokwebhayoloji (*it is that of biology*), ibhaysikili (*bicycle*), zebhayotheknoloji (*of biotechnology*), ibayoloji (*biology*)

The results above all include nouns which at least display noun class prefixes. In some cases these are preceded by other prefixes such as the copulative morpheme yi- as in yibhayotheknoloji (*it is biotechnology*), and a possessive morpheme as in zebhayoloji (*of biological*). The majority of identified nouns belong to the same semantic context as the input word.

For further illustration, the noun imibuzo (*questions*) occuring in Figure 3 is a sample entry that is partially enriched with the words ukubuza (*interrogation/to ask*) and ukuphendula (*to reply*), which are both semantically related lexical items and were also extracted based on word embeddings.

9 https://corpora.uni-leipzig.de/en?corpusId=zul_mixed_2019

4.3 Handling Faulty Input

Faulty or misspelled input words, or even out-of-vocabulary words, are a major usability issue. Users expect even for an "invalid" input to return a meaningful result, so methods for handling those use-cases are necessary.

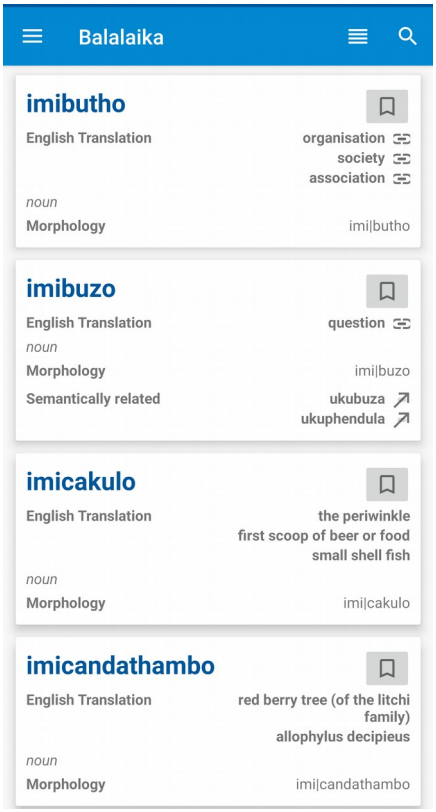

Figure 3: Xhosa dictionary entries partially enriched with references to semantically related terms (based on word embeddings)

Simple n-gram based methods can be used for searching for probable candidate words and suggesting those. A more comprehensive method is Fasttext (or more resilient word embeddings by Piktus et al., 2019), as it handles broken input rather well by using "sub-words" to infer embeddings for misspelled or unknown input words to then retrieve similar known words.[10] Single word embeddings in Fasttext are comprised of embeddings of variable length word n-grams and robust against slight changes in letters and work best with morphologically rich languages.[11] Prefixes and suffixes are more general in meaning due to their occurrence in many words in different contexts, stems however are more integral for the meaning as can be seen in the example above. That does not exclude semantically related words with completely different n-grams but those are ranked lower and additional post-processing may be necessary to only retrieve those words.

10https://fasttext.cc/docs/en/unsupervised-tutorial.html#importance-of-character-n-grams

11https://fasttext.cc/blog/2016/08/18/blog-post.html#works-on-many-languages

5. Conclusion

The sketched application is still under heavy development and therefore subject to changes. Its current state can already be examined at its public code repository; more extensive documentation about deployment or data import will be provided soon. A first feature-complete version can be expected by May 2020 and will incorporate the aforementioned data sets. In parallel, more approaches for improved access to and presentation of lexical data with a focus on less-resourced languages will be evaluated; suitable candidates will be implemented at a later stage.

6. Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017. Available: https://www.aclweb.org/anthology/Q17-1010, https://arxiv.org/abs/1607.04606

Bosch, S., Eckart, T., Klimek, T., Goldhahn, D., and Quasthoff, U. (2018). Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

Bosch, S. and Griesel, M. (2017). Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator* 38(1). http://www.literator.org.za/index.php/literator/article/view/1351

Büchler, M., Eckart, T., Franzini, G., and Franzini, E. (2017). Mining and Analysing One Billion Requests to Linguistic Services. In: Proceedings of The IEEE International Conference on Big Data 2016 (IEEE BigData 2016), Washington DC, 2016, 5-8. DOI: 10.1109/BigData.2016.7840979

Deng, Q. and Trainin, G. (2015). Learning Vocabulary with Apps: From Theory to Practice. *The Nebraska Educator: A Student-Led Journal*. 29. https://digitalcommons.unl.edu/nebeducator/29

Goldhahn, D., Eckart, T., Gloning, T., Dreßler, K., and Heyer, G. (2015). Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case. In: Proceedings of CLARIN Annual Conference 2015, Wrocław, Poland.

Goldhahn, D., Eckart, T., and Bosch, S. (2019). Enriching Lexicographical Data for Lesser Resourced Languages: A Use Case. In: Proceedings of CLARIN Annual Conference 2019. Eds. K. Simov and M. Eskevich. Leipzig, Germany.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Hinrichs, E. and Krauwer, S. (2014): The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

Kilgarriff, A. (2001). Comparing corpora. International journal of corpus linguistics, 6(1), 97-133.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). FastText.zip: Compressing text classification models. arXiv:1612.0365

Klimek, B. (2017). Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets.

Ordan, N. and Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation*, special issue on Lexical Resources for Machine Translation, 19(1):39–58. http://cs.haifa.ac.il/~shuly/publications/wordnet.pdf

Piktus, A., Edizel, N.B., Bojanowski, P., Grave, E., Ferreira, R., and Silvestri, F. (2019). Misspelling Oblivious Word Embeddings. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 3226–3234. https://www.aclweb.org/anthology/N19-1326/

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR, 2013.

Shu, R. and Nakayama, H. (2017). Compressing Word Embeddings via Deep Compositional Code Learning. http://arxiv.org/abs/1711.01068

Statistics South Africa. (2016). Community Survey. Pretoria: Statistics South Africa. http://cs2016.statssa.gov.za/wp-content/uploads/2016/07/NT-30-06-2016-RELEASE-for-CS-2016-_Statistical-releas_1-July-2016.pdf

Wilkinson M.D., Dumontier M., Aalbersberg, I.J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 160018.http://dx.doi.org/10.1038/sdata.2016.18, https://dash.harvard.edu/bitstream/handle/1/26860037/4792175.pdf