Exploring Afrikaans word embeddings with analogies and nearest neighbours

Gaustad, Tanja
Eiselen, Roald
Centre for Text Technology (CTexT), North-West University, South Africa
Tanja.Gaustad@nwu.ac.za
Roald.Eiselen@nwu.ac.za

Abstract

This paper presents an exploration of word embeddings for Afrikaans using the analogies and nearest neighbours methodologies. We compare the results on three types of embeddings (fastText, FLAIR and GloVe) on a novel analogy data set for Afrikaans, inspired by the Bigger Analogy Test Set: BATS (Gladkova *et al.* 2016). Our analysis shows that for Afrikaans, similar to English, the types of embeddings influence the quality of analogies found for different linguistic tasks. Our investigation also demonstrates, however, that these Afrikaans embeddings do not encode as clear a linguistic representation as with English embeddings. The exact reason for this is subject to future work, but the added morphological complexity and the lack of data most likely play a role.

Keywords: Text embeddings, Afrikaans, Analogy, Evaluation, Low-resource languages

1 Introduction and background

Over the last decade there has been a fundamental shift in the field of natural language processing (NLP) with the broad adoption of deep neural networks (DNNs), leading to major advances across the field. Underpinning this shift has been the introduction of more sophisticated methods for representing language data in numerical form, specifically vectorised real value representations known as word embeddings. These representations are a prerequisite for applying deep learning techniques to various NLP technologies. At the same time these embeddings have removed a significant portion of the linguistics that formed part of the NLP development cycle (even with traditional machine learning techniques) and resulted in a now almost completely engineering and state-of-the-art driven pursuit.

One of the features of these more complex representations is that there is no clear human interpretable connection between the vectorised representations and existing linguistic knowledge. This in turn makes the machine learning components, which are already very complex and difficult to interpret, almost impossible to fully understand. Even so, developers have made broad claims about the linguistic information that is represented in these embeddings on both morphological, syntactic, and semantic levels (Mikolov *et al.* 2013a, Pennington *et al.* 2014). To support these claims, different tests have been designed with the aim of indirectly explaining the information that is contained in the vector representations, primarily for English. More recently, there have also been more linguistically motivated investigations to attempt to get a better understanding of the information encoded in these embeddings and whether there are correlations with existing linguistic concepts and knowledge (Allen & Hospedales 2019, Miaschi & Dell'Orletta 2020, Warstadt *et al.* 2019).

For Afrikaans, there have been a limited number of investigations into the use of deep learning and word embeddings (Hanslo 2021, Heyns & Barnard 2020, Loubser & Puttkammer 2020, Ralethe 2020, Van Heerden & Bas 2021), mostly focussing on the application of deep learning to various NLP tasks. Until recently there were only three freely available Afrikaans word embedding models (Conneau *et al.* 2020, Grave *et al.* 2018), all without any direct assessment of their quality. Most recently, Eiselen (2022) released five new embedding models for Afrikaans (freely available from [1]), trained on a larger curated data set, of which three will be used in this study.

To our knowledge there has not been an in-depth investigation into the nature of word embeddings for Afrikaans, and whether the tests and claims made for English embeddings hold for a language such as Afrikaans, which is morphologically more complex both in terms of derivation and inflection, but also very productive in terms of compounding, unlike

English. Afrikaans also has substantially less data available to train these embedding models.

With this background in mind, our study aims at an exploratory investigation of Afrikaans word embeddings for three different architectures (GloVe, fastText, and FLAIR), applying existing evaluation techniques to answer the following questions:

- Do different embedding models encode different types of information for more morphologically complex and less resourced languages, such as Afrikaans?
- Are the intrinsic evaluation methods for English applicable to more morphologically complex languages, such as Afrikaans?

The following section provides a short overview of word embeddings, the three architectures under consideration and the training data used. Section 3 gives an overview of word embedding analysis techniques and the experimental design for Afrikaans, followed by an analysis of the results for the different experiments in Section 4. We conclude the investigation in Section 5 with further discussion of our findings and areas for possible future work.

2 Embedding architectures and training procedures

Finding meaningful numerical representations for text, and especially words, has a long history in NLP (Pennington *et al.* 2014). This is especially true in the machine learning context where these representations are a requirement for the models to be trained. Although work on learning these types of representations has been ongoing since Bengio *et al.* (2003), the predominant approach to representing words in machine learning models was so-called one hot vectors, where each word in a vocabulary is represented by a sparse vector containing zeros for all positions except the one for the particular word, which is set to 1. This method was usable, but only included information about whether the word is a member of the vocabulary or not. This changed in 2013 with the introduction of word2vec (Mikolov *et al.* 2013a, Mikolov *et al.* 2013b, Mikolov *et al.* 2013c), where real-valued vectors are learned by a combination of sentence level cooccurrences and a log-linear classifier to generate an output vector of predefined length. This was followed shortly thereafter by another embedding technique, Global Vectors (GloVe) (Pennington *et al.* 2014). Both methods allowed for training on huge amounts of data efficiently and the learned vector representations resulted in improvements in many downstream NLP technologies when combined with various deep learning techniques.

One of the major shortcomings of these "classic" embedding models is that each word has a single embedding, irrespective of the context in which the word appears. This has been addressed by more recent embedding and language models that leverage different DNN architectures, such as convolutional, recurrent, and transformer neural networks. These models learn a model for generating a vector output, which can adapt the vector representation for a word by taking the context in which the word appears into account. This has further allowed for major gains in downstream NLP tasks, at the cost of at least one very important aspect, namely explainability.

From the outset of developing embeddings, it was clear that although the vector representations did correlate with several semantic and morpho-syntactic attributes of English, it was difficult to determine what the model is learning. The nature of the embeddings - large vectors of real-valued numbers - and their training procedures obfuscate the meaning of a particular value in a particular vector position and how the values correlate with linguistic attributes. This has become even worse with the use of DNNs to generate the representations, since there are so many variables in the process, that it becomes almost impossible to determine if there are specific linguistic attributes associated with specific vector positions or regions. Even though there have been several efforts to propose methods for investigating embeddings, there is still no clear methodology for investigating the quality of the embeddings and explaining the values associated with the models. Furthermore, most of these investigations have focussed on English exclusively, and little work has been done to determine how representations perform in linguistically different and/or less-resourced

environments. For this study we concentrate on three embedding architectures, two of the most common classical embeddings, and one recurrent neural network, namely fastText, GloVe, and FLAIR embeddings.

fastText (Bojanowski *et al.* 2017) is an extension of the original word2vec (Mikolov *et al.* 2013a) that includes character n-grams in the embedding calculations to ensure that previously unseen words also generate embeddings. GloVe embeddings (Pennington *et al.* 2014) differ slightly from fastText in that they use global cooccurrences of words to train a log-bilinear regression model for generating the embeddings, and only consider words. Both of these models generate a single embedding for a word, irrespective of the context of the word. FLAIR embeddings on the other hand train a long-short-term-memory recurrent neural network to generate a representation based on a character sequence. This has two advantages: i) the same word in different contexts can have different representations reflecting the context; and ii) because the model considers characters, and not words, as basic units, any sequence of characters will get an embedding, irrespective of whether it has been seen during training. This last characteristic is especially useful in less-resourced environments where data sparsity remains a major issue. Both fastText and FLAIR each have two flavours, but due to space constraints we will focus only on the fastText continuous bag-of-words (CBoW) and FLAIR backward models in our analysis.

The primary prerequisite for training any type of embedding is a large collection of text data, typically in the order of billions of words. Unfortunately, no such large data collection exists for Afrikaans. For the purposes of this study, we used a combination of freely available data, including NCHLT Afrikaans Text Corpora (Eiselen & Puttkammer 2014), Autshumato Afrikaans monolingual text data (Snyman *et al.* 2013), and Wikipedia [2], as well as in-house data sets with restricted access due to copyright. In total, the models were trained on approximately 250 million words, which is far less data than is typically used in learning embeddings for most of the well-resourced languages of the world.

Since the current study is primarily interested in exploring the characteristics of the vector representations, default settings were used for training each of the embeddings.

3 Analysis techniques for word embeddings: Experimental design for Afrikaans

As mentioned above, a purely intrinsic evaluation of word embeddings remains elusive as the vectors contain large numbers of numeric values that do not clearly correspond to specific linguistic features and are therefore not easily interpretable by humans. Word embeddings are usually evaluated when used as input to a larger system which then shows improved performance. With this type of extrinsic evaluation it is difficult, however, to assess the input from the embeddings to the overall performance compared to e.g. the architecture of the system (Schnabel *et al.* 2015). We will now discuss how we used existing analysis techniques to evaluate and explore Afrikaans embeddings.

3.1 Analogies

There have been various attempts to investigate how embeddings for different words correlate and to show that they represent some (type of) linguistic attribute (Allen & Hospedales 2019, Miaschi & Dell'Orletta 2020, Tulkens *et al.* 2016, Warstadt *et al.* 2019). One such technique is to use analogy-based data to test the identification of linguistic relations using word embeddings (Mikolov *et al.* 2013a, Turney 2012). The most cited analogy is undoubtedly "Which word is to king as woman is to man?" with the expected answer "queen".

Mikolov *et al.* (2013a) introduced the Google analogy test set for English which contains nine morpho-syntactic and five semantic categories. The semantic tasks are all encyclopaedic whereas the morpho-syntactic categories include two tasks on derivational morphology, six on inflectional morphology and one encyclopaedic task, with between 20 and 70 unique word pairs each. As has been noted by Gladkova *et al.* (2016), there are two issues with existing test sets: firstly, most of them are not balanced for different types of linguistic relations and secondly, results are usually reported as an

average over an entire test set and not per type of relation. To remedy the first shortcoming, they introduced the Bigger Analogy Test Set (BATS) covering four main types of linguistic relations: inflectional and derivational morphology as well as lexicographic and encyclopaedic semantics. Each main type in turn contains 10 different relations with 50 unique word pairs each.

To date, there have been limited investigations of embeddings for South African languages (Dlamini *et al.* 2021), and no such analogy test sets exist for Afrikaans specifically. For this initial exploration of Afrikaans word embeddings, BATS served as inspiration to create a small set of analogies. We did not include any lexicographic semantic tasks at this stage but decided to focus on inflectional and derivational morphology plus two encyclopaedic semantics tasks for comparison with English.

The first step was a careful analysis of the categories used: being based on English, not all of them are applicable to a different language. For instance, one of the inflectional morphology tasks in BATS, verb plural formation, is not present in Afrikaans. Furthermore, for categories that are applicable, simple translation is usually not a viable option due to differences in usage, frequencies, and formations of words in Afrikaans. For each category covered in our study, we attempted to get a representative sample of as many aspects of the category as possible. For plural nouns for instance, a substantial number of different classes of regular and irregular plurals found in textbooks and grammars were included. The same holds for comparative and superlative adjectives. One linguistic aspect that has had limited investigation in this kind of testing, but is very prevalent in Afrikaans, is compounding. Therefore, a very small set of noun compounds was included to determine how they are represented in the embeddings.

Our test set for Afrikaans includes two semantic tasks, both encyclopaedic, and 11 morpho-syntactic tasks, three derivational, seven inflectional as well as compounding. Overall, there are 16,313 analogy "questions". Table 1 shows an overview of the categories chosen, including how many word pairs per task and an example for each.

Answers to analogy questions are calculated by taking the vector representation of word 1 (V_{w1}), subtracting the vector of word 2 (V_{w2}), related either semantically or morpho-syntactically, then adding the vector of a third word (V_{w3}). The resulting vector (V_{result}) is then compared to the vectors of all words in the model to find the vector(s) with the smallest Euclidean distance. The expectation is that the nearest vector to the result vector will express the same relationship to W3 as the relationship between W1 and W2. The prototypical example, $V_{king} - V_{man} + V_{woman}$ should result in a vector that has the smallest distance to V_{queen}. Similarly, $V_{stronger} - V_{strong} + V_{clear}$ should

Table 1: Analogy data set for Afrikaans: Types of linguistic relations, number of unique word pairs and examples.

Category	Subcategory	Task	# word pairs	Example
Morpho-syntactic	Derivational	Noun to Verb (*be-*, *ver-*)	20	man – beman taal – vertaal
		Noun to Adj (*-ies*)	10	simbool - simbolies
		Verb to Adj (*-baar*)	10	lees – leesbaar
	Inflectional	Adj comparative	41	duur – duurder
		Adj superlative	41	duur – duurste
		Adj comparative to superlative	41	duurder – duurste
		Attributive *-e*	10	teoreties – teoretiese
		Noun diminutive	56	hand – handjie
		Noun plural (reg/irreg)	76	kop – koppe
		Verb past tense	32	doen – gedoen
	Compounding	Noun compounding	19	landbousektor
Semantic	Encyclopaedic	Country - Capital	23	Duitsland – Berlyn
		Man - Woman	28	buurman – buurvrou

result in a vector closest to $V_{clearer}$. The vector visualisation in Figure 1 provides an intuition for why this should work. The offsets between man and woman, and king and queen, although not exactly the same, are similar. Therefore, removing the man characteristics from king, and adding woman's characteristics, should yield a vector in close proximity to queen.

3.2 Nearest neighbours

A second method described by Collobert & Weston (2008) and also referenced by Mikolov *et al.* (2013a) is the analysis of the nearest neighbours for a specific set of words. The nearest neighbour of a word is again determined by finding those word vectors which have the smallest Euclidean distance between the vector for the query and vectors for any other words for which embeddings exist in the model. The hypothesis is that a qualitative review of the neighbours provides additional insight into the types of relationships that the embeddings are learning. As an example, the fastText English embedding for the word "run" includes "runs, running, ran" which indicates the encoding of some morpho-syntactic properties. Although it is not possible to create a single metric for evaluation purposes, it is a useful procedure to gain an understanding of the underlying information that is encoded in the embeddings, such as hypernymy, hyponymy, synonymy, or morpho-syntactic relations.

One of the caveats to keep in mind with the nearest neighbour analysis is that different types of relations may be found within a single

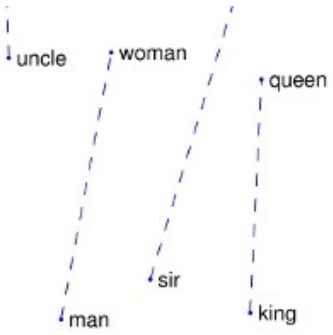

Figure 1: Gender related vector representation (from Pennington *et al.* (2014))

embedding architecture, and it may not always be immediately obvious what information is encoded in the embeddings. Consistent patterns can be found but should only be used to draw very broad and general conclusions about the encoded information.

To investigate the information encoded in the different embedding architectures for Afrikaans, we selected two words from each category listed in Table 1, 26 in total, and generated the five nearest neighbours for each word in each of the different architectures. These were then manually reviewed to determine the quality and nature of the embeddings.

3.3 Downstream task evaluation

The most common method for validating the quality of word embeddings is their application as part of a downstream task, such as POS tagging, named entity recognition or question answering. The use of embeddings rather than one hot encoding was one of the first steps enabling the current deep learning trend in NLP, and it has been consistently shown that using embeddings in downstream tasks improves the quality of the technology. This has also been shown to be the case for Afrikaans where a combination of FLAIR embeddings improve both POS tagging and NER results over previous models (Eiselen 2022). Due to space constraints we do not include these results in the current analysis.

4 Analysing Afrikaans word embeddings: Results

One of our initial motivations for this research was to investigate whether the type of information encoded in different embedding models is similar for linguistically different languages and whether methods used to evaluate English embeddings are also applicable to morphologically more complex languages. We will first present the results for the analogy task per linguistic (sub)task, including a thorough discussion of our observations. This intrinsic quantitative evaluation of the word embeddings for Afrikaans is followed by an intrinsic qualitative analysis investigating the nearest neighbours as described in section 3.2.

4.1 Analogies: Quantitative evaluation

Using the analogy data set for Afrikaans described earlier, the accuracy for each type of task is calculated separately. Previous studies mostly report only the accuracy of the word with the smallest Euclidean distance. Our evaluation, however, includes two accuracy scores: one for matches from the closest word (position 1) and one for matches from words in the subsequent four positions (position 2-5). Including more than only the closest matches gives us more insight into the different embedding representations for the various analogy tasks. Furthermore, when calculating accuracies, all input question words were excluded from the results. Omitting this adaptation resulted in much worse results (an effect also noted in Linzen (2016)). Table 2 shows the results for the different linguistic categories (aggregated at subcategory level) and types of embeddings.

In our experiments for Afrikaans, the overall best performing task and embedding type is GloVe on the semantic tasks with 51,11% accuracy, whereas the worst results are also obtained with GloVe, but on the derivational morphology tasks (1,11%). Compared to results for English on the Google data set (ranging from nearly 60% (Mikolov *et al.* 2013a) to high 60% (Levy & Goldberg 2014)), it is noteworthy how poorly all the embedding types perform on all of the tasks for Afrikaans. Contrasting our GloVe outcomes with Gladkova *et al.*'s (2016) more detailed results on BATS, the performance on Afrikaans is again quite a bit lower.

Focussing on the type of tasks, for the derivational tasks FLAIR performs best and GloVe worst. Both GloVe and fastText embeddings have a very high percentage of words not found in the top ten (87% and 68% respectively) which explains their poorer performance. The likely reason for the poor performance on derivations is the fact that per definition the paired words belong to different syntactic categories and typically do not appear in similar positions, hence do not have similar co-occurrences to the query word and will therefore have substantially different vector representations.

For inflectional morphology, fastText has the highest percentage of correct analogies for the first position, while GloVe has the lowest, although the difference in accuracy is fairly small compared to the other tasks. Interestingly enough, there are marked differences in correct words found in positions 2-5: FLAIR finds the searched for analogy in more than 40% of the cases, whereas the other embedding types only find it in slightly more than 20%. The FLAIR embeddings also find most analogies whereas GloVe finds the least. This can be explained by the fact that FLAIR embeddings encode character sequences and typically inflectional morphology happens at the character level. With regard to the subtasks for inflection, plural and diminutive forms are hardest to detect.

The results for the compound analogies indicate that the word embeddings do not learn a

Table 2: Accuracy for the Afrikaans analogy test set on different linguistic tasks (aggregated at subcategory level) for three word embedding types (best performance in position 1 per task type in bold).

Task Type	Architecture	Position 1	Positions 2-5	Not found
Derivational	fastText	11,94%	14,44%	68,06%
Derivational	FLAIR	**25,28%**	33,33%	33,06%
Derivational	GloVe	1,11%	9,44%	86,94%
Inflectional	fastText	**26,56%**	22,82%	41,51%
Inflectional	FLAIR	22,89%	41,45%	25,69%
Inflectional	GloVe	22,32%	20,98%	52,37%
Compounds	fastText	0,00%	0,00%	94,74%
Compounds	FLAIR	0,00%	0,00%	89,47%
Compounds	GloVe	0,00%	0,00%	100,00%
Semantic	fastText	16,56%	25,83%	48,10%
Semantic	FLAIR	7,92%	14,42%	72,58%
Semantic	GloVe	**51,11%**	33,20%	12,76%

representation of the constituents of the compound. Performing an analogy test that isolates the head of the compound results in a representation that is in a completely unrelated vector space, with no correlation to either the compound or its head. Although compounds are less frequent than the compound head in general, this does not seem to be the main contributing factor to the poor performance. As is discussed in the following section, the nearest neighbours of the head do contain many compounds, even relatively low frequency compounds, indicating that the full compound is seen as similar to the head, but not necessarily on a constituent level. This aspect of embeddings has not been studied extensively and will require further investigation in future.

The results for the semantic analogy tasks are the reverse of the morpho-syntactic ones (excluding compounding): GloVe very clearly outperforms all the other embedding types on all measurements. Here, the FLAIR embeddings perform the worst, also on all accounts. The one caveat to these results is that FLAIR embeddings are by nature contextual, and different vector representations will be generated when considering the words in sentence contexts, which was not the case in our tests. It may well be that the FLAIR embeddings perform better on semantic analogy tasks if vectors are generated for words in a sentence context. Unfortunately, there is not currently a well-defined methodology for generating embeddings for this type of task and it is something that will need to be considered in future work, especially if this type of analysis is undertaken with other types of representations, such as transformer models.

To summarize, our results corroborate earlier findings on English that different types of embeddings work best for different linguistic analogy tasks. In addition, our results on this analogy test set indicate that inflectional morphology is easier to model than derivational morphology, whereas compounding, a typical feature of Afrikaans, is very difficult to model. Overall, performance on Afrikaans, a more morphologically complex and productive language, is poorer than expected.

4.2 Nearest neighbours: Qualitative evaluation

After the more quantitative analysis using analogies, we now examine the nearest neighbours for Afrikaans, whether they differ from our expectations, and what we can learn from this examination.

fastText embeddings

The main finding for the fastText embeddings is that there are little to no examples of semantic relations in neighbours for any of the words selected, and in almost all cases the query is a substring within the set of nearest neighbours, see e.g. for *verdeel* (divide) and *Berlyn* (Berlin):

> *verdeel* ▭ *opverdeel, onderverdeel, onverdeel, verdeelpyp, verdeelbaar*
>
> *Berlyn* ▭ *Berlyn-Schönefeld, Berlyner, Berlynse, Berlynmuur, Wes-Berlyn.*

This can primarily be attributed to the fact that the inclusion of subword information in the embeddings has a strong effect on the vector representations and coincides with the fact that the morphology of Afrikaans is more productive than English, both in terms of inflectional and derivational paradigms. The consequence of this is that any inflectional or derivational form exhibiting some form of typographic change, e.g. shortening of the double vowels in plural forms, are not typically associated with the query word and therefore not returned as nearest neighbour. Furthermore, Afrikaans being a compounding language means that a large number of words closely associated with a query tend to be either inflections of the query or a compound including the query, rather than semantically related words as is often the case in English.

FLAIR embeddings

As was previously shown in Section 4.1, and expected given the evaluation parameters, there are essentially no semantic relationships between the query words and nearest neighbours for the FLAIR embeddings. Unlike the fastText embeddings, the FLAIR embeddings do not include the query term as a substring of the neighbours, but there are strong correlations with inflectional patterns. As an example, the nearest

neighbours for the word *leesbaar* (readable) are as follows:

> *leesbaar* ▭ *leeservaring, leefwêreld, kwesbaar, leefwyse, leesstof, leefstyl, vloeibaar, voorspelbaar, aanpasbaar*

From this set we see that the model either agrees with the 'lee' substring at the beginning of the word or the *-baar* (-able) morpheme at the end. This is an indication that the model is more likely to model affix structure.

GloVe embeddings

The embeddings for GloVe are substantially different from the other types, with a combination of morpho-syntactic, semantic, and cooccurrence instances showing up in the list of nearest neighbours, for example:

> *hoog* ▭ *hoë, laag, bo, bokant, hoogte, hoër, ver, meter, so*

> *ironie* ▭ *humor, satire, sarkasme, simboliek, tikkie, ironiese, tragiese*

In the examples for *hoog* (high), there are inflections - *hoë* (high), *hoogte* (height), *hoër* (higher); semantically related words - *laag* (low), *bo* (above), *bokant* (above, top); as well as words that frequently cooccur with *hoog* - *meter hoog* (meter's high), *so hoog* (so high). These cooccurences are not necessarily the most frequent as *te* (too), *is* (is) and *baie* (very) all occur more frequently with *hoog* than *meter* (VivA 2022). The same pattern also occurs for *ironie* (irony) with all three types of relations found in the nearest neighbours.

5 Discussion and future work

Our explorations of Afrikaans embeddings have shown that, similar to other languages, different types of embeddings work best for different linguistic analogy tasks. However, a careful analysis of the analogies and nearest neighbours results also demonstrates that these embeddings do not encode as clear a linguistic representation as for English. There are two possible reasons for these differences: Afrikaans is linguistically different to a relevant degree or more data is needed to train more representative embeddings. Currently, we do not know what the main source of the shortcomings for Afrikaans embeddings is, but surmise that most likely both the added morphological complexity and the lack of data have an influence.

In the case of linguistic diversity/complexity, this would mean the more different a language is compared to English, e.g. other South African languages such as isiZulu or Setswana, the more carefully word embeddings should be trained and the more critically they have to be evaluated. If data sparsity is the culprit (even though 250 million is middle-ground in terms of resources), this does not bode well for resource-scarce languages when building and subsequently using embeddings for NLP tasks and/or trying to understand what they represent.

Overall, as embeddings for morphologically complex, compounding languages are substantially different to English, both how to train these embeddings as well as how we analyse them need to be rethought, especially for under-resourced languages.

Future work includes building a full analogy set covering more linguistic categories relevant for Afrikaans. Expanding these explorations to the other South African languages is also an interesting challenge, especially given their high morphological productivity in conjunction with very little data.

Notes

[1] https://repo.sadilar.org

[2] https://dumps.wikimedia.org/

Acknowledgements

This work was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Sports, Arts and Culture.

References

Allen, C & Hospedales, T 2019, 'Analogies explained: Towards understanding word embeddings', *In*: Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, PMLR, pp. 223-231.

Bengio, Y, Ducharme, R, Vincent, P & Jauvin, C 2003, 'A neural probabilistic language model',

Journal of Machine Learning Research, vol. 3**,** pp. 1137–1155.

Bojanowski, P, Grave, E, Joulin, A & Mikolov, T 2017, 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics,* vol. 5**,** pp. 135-146.

Collobert, R & Weston, J 2008, 'A unified architecture for natural language processing: Deep neural networks with multitask learning', *In*: Proceedings of the 25th International Conference on Machine learning, Valencia, Spain, pp. 160-167.

Conneau, A, Khandelwal, K, Goyal, N, Chaudhary, V, Wenzek, G, Guzmán, F, Grave, É, Ott, M, Zettlemoyer, L & Stoyanov, V 2020, 'Unsupervised Cross-lingual Representation Learning at Scale', *In:* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, ACL, pp. 8440-8451.

Dlamini, S, Jembere, E, Pillay, A & van Niekerk, B 2021, 'isiZulu Word Embeddings', *In:* Proceedings of the 2021 Conference on Information Communications Technology and Society, Durban, South Africa, IEEE, pp. 121-126.

Eiselen, R 2022, 'Afrikaans Text Embeddings for Sequence Labelling with Deep Neural Networks', *In:* Proceedings of the Southern African Conference for Artificial Intelligence Research 2022, Stellenbosch, South Africa, SACAIR.

Eiselen, R & Puttkammer, MJ 2014, 'Developing Text Resources for Ten South African Languages', *In:* Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, ELRA, pp. 3698–3703.

Gladkova, A, Drozd, A & Matsuoka, S 2016, 'Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't', *In:* Proceedings of the NAACL Student Research Workshop, San Diego, CA, ACL, pp. 8-15.

Grave, É, Bojanowski, P, Gupta, P, Joulin, A & Mikolov, T 2018, 'Learning Word Vectors for 157 Languages', *In:* Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, ELRA, pp 3483-3487.

Hanslo, R 2021, 'Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages', *In:* Proceedings of the 16th Conference on Computer Science and Intelligence Systems (FedCSIS), Virtual, IEEE, pp. 115-119.

Heyns, N & Barnard, E 2020, 'Optimising word embeddings for recognised multilingual speech', *In:* Proceedings of the Southern African Conference for Artificial Intelligence Research Conference 2020, Online, Virtual, SACAIR, pp. 102-116.

Levy, O & Goldberg, Y 2014, 'Linguistic Regularities in Sparse and Explicit Word Representations', *In:* Proceedings of the 18th Conference on Computational Language Learning, Baltimore, MD, ACL, pp. 171-180.

Linzen, T 2016, 'Issues in evaluating semantic spaces using word analogies', *In:* Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany, ACL, pp. 13-18.

Loubser, M & Puttkammer, MJ 2020, 'Viability of Neural Networks for Core Technologies for Resource-Scarce Languages', *Information,* vol. 11**,** pp. 41-57.

Miaschi, A & Dell'Orletta, F 2020, 'Contextual and non-contextual word embeddings: an in-depth linguistic investigation', *In:* Proceedings of the 5th Workshop on Representation Learning for NLP, Seattle, WA, ACL, pp. 110-119.

Mikolov, T, Chen, K, Corrado, G & Dean, J 2013a, 'Efficient estimation of word representations in vector space', *In:* Proceedings of the International Conference on Learning Representations 2013, Scottsdale, AZ.

Mikolov, T, Sutskever, I, Chen, K, Corrado, GS & Dean, J 2013b, 'Distributed representations of words and phrases and their compositionality', *Advances in Neural Information Processing Systems,* vol. 26**,** pp. 3111–3119.

Mikolov, T, Yih, W & Zweig, G 2013c, 'Linguistic regularities in continuous space word

representations', *In:* Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, ACL, pp. 746–751.

Pennington, J, Socher, R & Manning, CD 2014, 'Glove: Global vectors for word representation', *In:* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, ACL, pp. 1532–1543.

Puttkammer, MJ, Eiselen, R, Hocking, J & Koen, F 2018, 'NLP Web Services for Resource-Scarce Languages', *In:* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Melbourne, Australia, ACL, pp. 43–49.

Ralethe, S 2020, 'Adaptation of deep bidirectional transformers for Afrikaans language', *In:* Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, ELRA, pp. 2475-2478.

Schnabel, T, Labutov, I, Mimno, D & Joachims, T 2015, 'Evaluation methods for unsupervised word embeddings', *In:* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, ACL, pp. 298-307.

Snyman, DP, McKellar, CA & Groenewald, H 2013, 'Autshumato English-Afrikaans Parallel Corpora', Data set, v1.0, South African Centre for Digital Language Resources (SADiLaR), https://hdl.handle.net/20.500.12185/397.

Tulkens, S, Emmery, C & Daelemans, W 2016, 'Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource', *In:* Proceedings of the 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia, ELRA, pp. 4130-4136

Turney, PD 2012, 'Domain and function: A dual-space model of semantic relations and compositions', *Journal of Artificial Intelligence Research,* vol. 44, pp. 533-585.

Van Heerden, I & Bas, A 2021, 'AfriKI: Machine-in-the-Loop Afrikaans Poetry Generation', *In:* Proceedings of the 1st Workshop on Bridging Human-Computer Interaction and Natural Language Processing, Virtual, ACL, pp. 74-80.

VivA 2022, 'Korpusportaal: Omvattend 1.11', *Virtuele Instituut vir Afrikaans (VivA)*, viewed 1 September 2022, <https://viva-afrikaans.org>.

Warstadt, A, Cao, Y, Grosu, I, Peng, W, Blix, H, Nie, Y, Alsop, A, Bordia, S, Liu, H & Parrish, A 2019, 'Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs', *In:* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, ACL, pp. 2877-2887.