Analysing Afrikaans lexical blends using Levenshtein distances

Benito Trollip (South African Centre for Digital Language Resources (SADiLaR), North-West University, Potchefstroom, South Africa)
benito.trollip@nwu.ac.za

Trudie Strauss (Department of Afrikaans and Dutch, German and French, University of the Free State, Bloemfontein, South Africa)
strausst@ufs.ac.za

Abstract

The utility of language is not limited to its communicative function as can be illustrated by constructions like *hangry:* Two words (*hungry* and *angry*) are combined to generate a new construction that describes a state of being angry due to being hungry. These constructions are known as lexical blends. Language users can create blends for purposes ranging from literary effect to displaying linguistic creativity.

In this paper Afrikaans blends (e.g., *kapoen* as a blend of *kak* 'shit' and *pampoen* 'pumpkin') are investigated. Context is given with reference to available studies before the analysis of a dataset of Afrikaans blends is undertaken. The collected data is analysed using the Levenshtein distance metric, a type of edit distance that measures the similarity between two strings in terms of the number of single-character edits to illustrate similarity between source words and blends. The following hypothesis is investigated: Whether the shorter source word in a blend contributes more to the blend. From the available data we cannot confirm a positive tendency toward this hypothesis and argue that we require more data before any kind of conclusion can be drawn. Still, this study shows to what degree edit distance measuring can be employed to lay the foundation for the description of Afrikaans blends.

Keywords: Afrikaans, lexical blend, Levenshtein edit distance, morphology, portmanteau, recognisability, reductive, similarity, source word length, word-formation

Opsomming

Analise van Afrikaanse versmeltings met behulp van Levenshteinafstande

Die nut van taal is nie tot die kommunikatiewe funksie daarvan beperk nie, soos geïllustreer kan word deur konstruksies soos *hangry*: Twee woorde (*hungry* en *angry*) word gekombineer om 'n nuwe konstruksie te vorm wat 'n toestand van woede as gevolg van hongerte beskryf. Hierdie konstruksies staan as versmeltings bekend. Taalgebruikers kan versmeltings skep vir doeleindes wat wissel van literêre effek tot om hul linguistiese kreatiwiteit ten toon te stel.

In hierdie artikel word Afrikaanse versmeltings (bv. *kapoen* as 'n versmelting van *kak* en *pampoen*) ondersoek. Konteks word eerstens met verwysing na beskikbare studies verskaf, alvorens 'n datastel van Afrikaanse versmeltings ontleed word. Die versamelde data word ontleed deur gebruik te maak van die Levenshtein-afstandsmetriek, 'n tipe wysigingsafstand,

How to cite this article:
Trollip, Benito & Trudie Strauss. 2023. "Analysing Afrikaans lexical blends using Levenshtein distances". In *Proceedings of the 4th International Afrikaans Grammar Workshop,* edited by Adri Breed. Potchefstroom: North-West University. pp. 1-16. DOI: 10.25388/nwu.25052690.

wat die ooreenkomste tussen twee karakterstringe met verwysing na die aantal enkelkarakter wysigings tussen bronwoorde en versmeltings meet. Die hipotese wat ondersoek word, is of die korter bronwoord van ’n versmelting meer van sy vorm tot die versmelting bydra. Uit die beskikbare data kan ons nie ’n positiewe tendens ten gunste van die hipotese bevestig nie. Ons voer aan dat meer data benodig word om enige gevolgtrekkings oor die hipotese te kan maak. Desnieteenstaande word daar in hierdie artikel aangetoon tot watter mate die meting van wysigingsafstande die basis verskaf vir die beskrywing van versmeltings in Afrikaans.

Sleutelwoorde: Afrikaans, bronwoordlengte, herkenbaarheid, Levenshteinwysingsafstand, morfologie, portmanteau, reduksie, soortgelykheid, versmelting, woordvorming

1 Introduction

Besides language’s communicative function, it also serves as an important medium for creative expression (Carter 2004; Gries 2004a; Monakhov 2021). Even though language is communicative irrespective of how creative some constructions could be perceived, the aim of some constructions is less on conveying literal meanings or speech acts and more on displaying linguistic creativity. An example of such a construction is *ka|poen*[1] where two autonomous words (*kak* ‘shit’ and ~~pam~~poen ‘pumpkin’) are blended into a new construction (or form-meaning pair) that describes a colour deemed (negatively) to be a mixture or combination of the colours of shit and pumpkin. From this example it is possible to identify that *kak* contributes two of its three graphemes to the blend, while *pampoen* contributes four of its seven graphemes. Constructions like *ka|poen* where words are combined in a way that combines part of the forms (in the form of graphemes and/or phonemes) and meanings of two words to form a new construction, are known as lexical blends (Gries 2004b).

The content of this article includes an overview of existing research on lexical blends in section 2 to give an impression of the current state of the art. Currently there is a critical mass of literature describing various aspects of blends specifically in English (Bauer 2012, 11). Contrary to the situation in English, lexical blends in Afrikaans have not been described or studied in any detail yet. To fill this descriptive gap, the focus in this article is on describing lexical blends in Afrikaans from a usage-based perspective, accepting the description of Kemmer (2003). The statement by Kemmer (2003, 91), confirming the aptness of describing blends through a schema-based theory (specifically cognitive grammar), motivates the use of her theoretical description as the basis of usage-based approach in this article. Our focus is restricted to the analysis of the realised forms (i.e., graphemes used to represent the construction) of blends, not necessarily of the way meaning is constructed.

In section 3 the compilation and annotation of observed Afrikaans blends, following the annotations used in Wulff and Gries (2019), will be elaborated on. The annotated blend dataset is a resource developed for purposes of this article and will be discussed in section 3 in more detail. The collected data is analysed using the Levenshtein (1966) distance metric, a type of edit distance that measures the similarity between two strings with reference to the

[1] The writing conventions used in Wulff & Gries (2019) will be followed in this paper. The | (pipe) symbol is used to indicate blend boundaries. The boundary will be indicated where it is possible to discern where the graphemes from the first source word end and where those used from the second source word begin. Strikethroughs are used to indicate which parts of the source words are not included in the blend.

number of single-character edits to illustrate similarity between source words and blends. Through this we aim to show whether one of the hypotheses formulated by Wulff and Gries (2019) also holds for Afrikaans blends, namely that the shorter source word contributes more to the blend than the longer source word.

The aim of this paper is therefore not to provide a comprehensive description of Afrikaans lexical blends, but to evaluate to what degree the specified hypothesis, which Wulff and Gries (2019) have shown to hold for English, is also relevant for Afrikaans. The implications of using a small data sample, like the Afrikaans blend dataset, as input for calculating edit distances are also discussed. This article concludes with section 4 where possibilities for further research, focusing on how Afrikaans blends could be handled in Afrikaans linguistic research, will also be given.

2 Literature review

An overview of existing literature on lexical blends is required to provide context about the main challenges when studying blends. This section is divided into three parts: section 2.1 includes a general overview of the literature; section 2.2 deals with studies of blends in languages other than English; section 2.3 concludes section 2 with a reflection on the nature of blends.

2.1 General overview

Blends have been a topic of interest to linguists for decades. Existing literature on lexical blends includes older sources ranging from Cannon (1986), Kelly (1998) and Kubozono (1990) to more recent publications like Baliaeva (2016; 2022), Bauer (2006; 2012) and Gries (2004a; 2004b; 2006; 2012). The continued interest in these constructions is further illustrated by the recent thesis by Kjellander (2022). These studies focus on various aspects of blends – if the work of Gries is considered, there is a definitive focus on systematising the structure of blends. In his most recent sole authored work on blends Gries (2012) places more focus on the cognitive and psycholinguistic aspects of blends. In his recommendation for future studies on blends, he recommends that purely descriptive work should not be undertaken without consideration of psycholinguistic concepts (Gries 2012, 166). The structure of blends and, more specifically, how the source words' structures are combined to form the blend and the meanings combined to form the blended meaning are generally emphasised.

Gries (2006, 537-38) mentions three trends in lexical blend literature; studies classifying blends or differentiating them from other word-forming processes like affixation, studies focused on how blends are formed to still be identifiable with their source words, and studies that are not convinced that there are discernible patterns or an order to blend formation. More recently Bauer (2012) considered existing literature on (English) blends and discusses ordering, recognisability, formal expectations, as well as the semantics of blends as central aspects. Bauer (2012, 21) states that it is important to delineate blends clearly and to do away with how "fuzzy" the constructions are.

Despite the vast amount of literature on blends, uncertainty about which constructions really count as blends and which do not, persists (Kjellander 2022, 21). Despite this uncertainty, Gries (2012, 146) formulates a definition of blends a decade before Kjellander's study.

According to Gries a blend is “an intentional fusion of typically two (but potentially more) words where a part of a first source word (sw1) – usually this part includes the beginning of sw1 – is combined with a part of a second source word (sw2) – usually this part includes the end of sw2 – where at least one source word is shortened and/or the fusion may involve overlap of sw1 and sw2.” For the purposes of this article this definition will be utilised, and the data will only include blends of two words.

Constructions sometimes discussed in conjunction with blends are known as blend splinters (Jurado 2019; Lalić–Krstin, Silaški, and Đurović 2023). Examples of blend splinters include *-gasm* and *-holic* like in the examples *nerdgasm*, *musicgasm*, *shopoholic* and *tweet-aholic*. A marked difference therefore between splinter constructions and blends is the productive forming of new constructions using a specific and more affix-like component. These constructions are therefore mentioned for purposes of completeness, but they will not be considered further.

There is another reductive word-forming process that should also be distinguished from blends, namely clipped compounds. Bauer (2006, 501-03) distinguishes between clipped compounds where the first parts of two words are clipped and combined (e.g., *sitcom* from *situation* and *comedy*) and ‘real’ blends where the first part of one source word and the second part of a second source word are combined (e.g., *monergy* from mon~~ey~~ and ~~en~~ergy). Combrink (1990) makes the same distinction for the purpose of Afrikaans but does not elaborate more than that. These constructions could qualify as blends according to Gries’ definition due to him using “usually” when specifying that the first part of the first source word and second part of the second source word is combined to form blends. For our purposes, the interpretation of blends will be in the stricter sense as stated by Bauer (2006) and Combrink (1990).

2.2 Blends in other languages

Research either focusing on blends in languages other than English is less abundant. Most of the literature that either considers blends comparatively to English or that exclusively describes blends in other languages, is recent compared to the works discussed in the previous section. It is therefore an innovative perspective to consider these constructions in other, less-studied or less-described languages. In this respect Kjellander (2022, 2) highlights two aspects that could be influencing the persisting gap in blend research, namely blending being regarded as “an insignificant linguistic phenomenon” and “the challenging technical problems of localizing [blends]”. It will therefore actively aid the study of lexical blends to describe them in languages other than English.

As a starting point to considering studies not solely focused on English, there are studies in which English blends are compared to blends in other languages. Comparative studies have been undertaken that consider blends in French and English (Renner 2019), Spanish and English (Balteiro 2018), and even Arabic and English (Mohsin 2020). It is therefore an emerging trend to use English as a base and compare it to more related languages like French and completely unrelated languages like Arabic. Renner (2015) goes further and includes blends from German, Basque, Serbian and Spanish, amongst others. His focus is on the aspect of wordplay that blends represent. He concludes that the creation of new blends constitutes an act of wordplay, and that the use and construction of blends could promote

social interaction (Renner 2015, 130-1). There are language specific studies focused on Italian (Cacchiani 2016; Micheli 2022), Ukrainian (Winters 2017) and Portuguese (Villalva and Minussi 2022). A wide range of languages and language groups are represented in recent work, but blend research in West-Germanic languages, a subfamily that includes Afrikaans, seem to be relatively absent.

Considering the limited amount of work about other languages, it should not be surprising that there are no studies dedicated to the study of blends in Afrikaans. Afrikaans blends have received cursory to no attention if one considers sources dedicated to Afrikaans morphology. In works dedicated to Afrikaans morphology, like Combrink (1990), he mentions blends in passing and distinguish between reduction compounds (discussed earlier as *clipped compounds*) versus 'real' blends, in line with Bauer (2012). Similarly, Van Huyssteen (2017) refers to Combrink and only makes a cursory remark on blends in his recent book chapter on Afrikaans morphology. The most comprehensive discussion on blends in Afrikaans is Van Huyssteen (2020). In the latter description blends are categorised as a form of subtractive word-formation in Afrikaans and numerous examples are given to the reader to illustrate the different forms blends could take. This article serves as a renewed attempt to illustrate how Afrikaans blends are a worthwhile topic of inquiry for linguists.

2.3 The nature of blends

Having already given an overview of research on lexical blends, whether it be theoretical or language specific descriptions, it remains to reflect on the nature of blends. When considering the nature of blends a particular focus on language users' creativity leads to the constructions being characterised by some linguists as unpredictable or unsystematic (Connolly 2013). A completely opposing perspective is that blends indeed are systematic (Gries 2006; 2012). The playful nature of blends is also generally accepted (Kjellander 2019; Renner 2015).

Besides creativity, systematicity, and playfulness, blends tend to be mentioned with reference to recognisability as well. To illustrate the different possible levels of recognisability of some Afrikaans blends, consider *web*|*inaar* 'a seminar on the web' (*web* and ~~*sem*~~*inaar*) and *sensi*|*teef* 'being sensitive causing one to be a bitch' (*sensi*~~*tief*~~ and *teef*). The first case is almost self-explanatory in terms of meaning, possibly owing to its similarity to and identifiability with the English *web*|*inar*, while the second needs more context before one could arrive at the correct meaning. The composite meanings of *sensitief* 'sensitive' and *teef* 'bitch' are utilised in such a way that a metaphorical meaning of *teef* is activated in the blend, rather than the literal meaning.

Wulff & Gries (2019) recently explored a more quantitative method to detect and analyse blends. An aim expressed in their work is to specifically improve the observational approach in lexical blend research. They therefore propose a more systematic method of investigation, building on the already quite exhaustive mass of observational data that has already been collected and described over the past decades. This is currently a challenge in the case of Afrikaans due to the complete lack of studies on lexical blends observational or otherwise. Therefore, the data used for this article is of an observational nature, as it marks the beginning of filling the descriptive gap.

3 Data collection and analysis

As stated in the introduction, the part of the analysis in Wulff and Gries (2019) where English blends are analysed using the Levenshtein distance metric serves as the basis for our analysis. Despite their critique toward the number of observational studies focusing on lexical blends, there is no denying that observational work has laid the foundation for more recent empirical or quantitative work. The blends in Wulff and Gries (2019) were sourced through asking participants to construct new (English) blends given specific source words. The experimental design of that study therefore differs from the current article in as far as no participants were involved in sourcing the Afrikaans blends. To calculate the edit distances usage data is needed. In the remainder of section 3 the data collection and analysis will be discussed.

3.1 Afrikaans blend data

Kjellander (2019) states that sourcing blends systematically is a challenge considering that they are, in his words, “ephemeral, informal, creative, and complex”. Considering that he states this with reference to English, it could serve as a deterrent for envisaged data-driven studies of blends in other languages. The examples in Van Huyssteen (2020) are the most extensive collection of Afrikaans blends that could be found. To enable our investigation, a dataset with usage examples is therefore needed.

Without available data generalisations or hypotheses about aspects of Afrikaans blends, completely replicating studies like that of Wulff and Gries (2019) are not feasible. The reason for that is that there is a lack of identified and described blends to perform quantitative work on. For this reason, the current Afrikaans lexical blend dataset (Trollip and Strauss 2023) contains instances of Afrikaans blends, originally gathered observationally and then queried in an online Afrikaans corpus.[2] The summary of constructions in Table 1 consists of all the blends observed. The blend is given in the first column, the first source word (SW1) in the second column and the second source word (SW2) in the third column. The list in Table 2 features the blends with at least one hit in either the Comprehensive corpus of the Virtual Institute for Afrikaans (VivA-CPC) or the Exclusive corpus (VivA-CPE) (VivA 2020a; 2020b). The first three columns of Table 1 and Table 2 are identical, with the addition of the frequencies from VivA-CPC and VivA-CPE in columns four and five, respectively. For the duration of the article the blends in Table 2 serve as the dataset.

Table 1: Observed Afrikaans lexical blends

Blend	Source word 1 (SW1)	Source word 2 (SW2)
atmosvuur	atmosfeer	vuur
betereinder	beter	bittereinder
bewarea	bewaar	area
dagmerrie	dag	nagmerrie
depresSeeff	depressief	Seeff

[2] In the final dataset the observed blends (Table 1) and the blends that had hits in VivA’s corpora (Table 2) are included.

Blend	Source word 1 (SW1)	Source word 2 (SW2)
Deurnis	deur	deernis
dikkelicious	dik	delicious
dramedie	drama	komedie
Englikaans	Engels	Afrikaans
Funky-kaans	Funky	Afrikaans
glitterati	glitter	literati
groenflasie	groen	inflasie
Hellington	hel	Wellington
hoentaal	hoender	tarentaal
hoeranje	hoer	oranje
hominee	homoseksueel	dominee
jimpel	jags	simpel
kapoen	kak	pampoen
Karoomaties	Karoo	idiomaties
Katarstrofe	Katar	katastrofe
knofsielie	knoffel	pietersielie
Kuberkulose	Kuber	tuberkulose
labrahond	labrador	hond
lam-brador	lam	labrado
lammetjie-uitnek	lammetjie	rammetjie-uitnek
(lank-)moerig	lank	moerig
Mengels	meng	Engels
midbyt	middagete	ontbyt
miljuisend	miljoen	duisend
mismoerig	mislik	moerig
monger	moerig	honger
onderweldig	onder	oorweldig
oueteparadys	ouetehuis	paradys
pikkewouter	pikkewyn	kabouter
(P)Oesjaar	poes	oesjaar
poeskantoor	poes	poskantoor
poeslisie	poes	polisie
ProteJA	Protea	ja
Ramaforie	Ramaphosa	euforie
Ramaphala	Ramaphosa	Phala-Phala
Ramaphorie	Ramaphosa	euforie
rib-bels	rib	rebels
romanteties	romanties	pateties
saterdrag	Saterdag	drag
semigrasie	semi	emigrasie
Sensiteef	sensitief	teef
sier-druiwe	sier	suurdruiwe
snoepsie	snoep	oepsie

Blend	Source word 1 (SW1)	Source word 2 (SW2)
stormkopies	storm	inkopies
Suip-Afrika	suip	Suid-Afrika
Twitterati	Twitter	literati
uitgemat	uitgeput	afgemat
verstoptimisme	verstop	optimisme
webinaar	web	seminaar
Zupta	Zuma	Gupta

The final dataset summarised in Table 2 includes 30 blends, ranging from highly frequent blends (like *web*|*inaar* and *Zu*|*pta*) to extremely infrequent blends (like *pikkew*|*outer* and *Rama*|*phala*). From the highly frequent ones there is *web*|*inaar that* has 173 hits in VivA-CPC, but only 2 hits in VivA-CPE and *Zu*|*pta* that has 72 hits in VivA-CPC and 431 hits in VivA-CPE. Many of the blends have extremely low frequencies – 10 of the 30 blends are hapax legomena (have only a single hit) in VivA-CPC. To put into context how low these frequencies are, one can consider that the version of VivA-CPC queried consists of 299 539 233 words and VivA-CPE of 64 508 637 words. It is possible that these specific blends have these low frequencies in the specific texts these corpora consist of or, more probably, that blends are still a very uncommon occurrence in written Afrikaans.

Table 2: Afrikaans lexical blends confirmed in VivA's corpora

Blend	Source word 1 (SW1)	Source word 2 (SW2)	Frequency in VivA-CPC 1.12	Frequency in VivA-CPE 1.13
webinaar	web	seminaar	173	2
bewarea	bewaar	area	169	5
Zupta	Zuma	Gupta	72	431
Ramaforie	Ramaphosa	euforie	40	18
Deurnis	deur	deernis	32	0
semigrasie	semi	emigrasie	9	2
dagmerrie	dag	nagmerrie	9	1
kapoen	kak	pampoen	8	8
dikkelicious	dik	delicious	8	0
Twitterati	Twitter	literati	7	6
Mengels	meng	Engels	6	1
jimpel	jags	simpel	5	3
betereinder	beter	bittereinder	5	0
glitterati	glitter	literati	5	0
poeslisie	poes	polisie	3	6
sensiteef	sensitief	teef	2	1
Englikaans	Engels	Afrikaans	2	0
hoeranje	hoer	oranje	2	0
Kuberkulose	kuber	tuberkulose	2	0
onderweldig	onder	oorweldig	2	0
poeskantoor	poes	poskantoor	1	3

Blend	Source word 1 (SW1)	Source word 2 (SW2)	Frequency in VivA-CPC 1.12	Frequency in VivA-CPE 1.13
Suip-Afrika	suip	Suid-Afrika	1	2
miljuisend	miljoend	duisend	1	1
(lank-)moerig	lankmoedig	moerig	1	0
dramedie	drama	komedie	1	0
midbyt	middagete	ontbyt	1	0
mismoerig	mismoedig	moerig	1	0
pikkewouter	pikkewyn	kabouter	1	0
saterdrag	Saterdag	drag	1	0
Ramaphala	Ramaphosa	Phala-Phala	0	1

3.2 Edit distance

Jurafsky and Martin (2008:74-7) discusses the application and uses of measuring the distances (or differences) between strings. It is particularly important to identifying word errors as is the case when developing spellcheckers, and edit distances are useful across different areas within speech and language processing. The intention with using the Levenshtein distances with the Afrikaans blend dataset is to show how different each blend is from its source words. The differences can include additions, deletions, and substitutions. As an example, consider *milj*|*uisend* that has *miljoen* 'million' as SW1 and *duisend* 'thousand' as SW2. To get from *miljoen* to *milj*|*uisend* requires that *-oen* be substituted with *-uis* (one distance measure per substitution) and *-end* be added (one distance measure per addition). The number of operations from *miljoen* to *milj*|*uisend* is therefore 6. When considering the change from *duisend* to *milj*|*uisend* it is required that *-d be replaced by milj-* (one distance measure per addition). The number of operations from *duisend* to *milj*|*uisend* is 4. Therefore, if only the number of operations is considered, *duisend* is closer to *milj*|*uisend* than *miljoen.* To take into consideration the fact that the words are not of the same length, this distance may further be relativised by dividing by the length of the blend, in the same way that Wulff and Gries (2019, 16) calculate the Levenshtein string-edit distance. In other words, the string edit distance from *miljoen* to *miljuisend would be* $\frac{6}{10}$, because the longer of the two words consists of 10 characters and as shown above one needs 6 operations to go from *miljoen* to *milj*|*uisend*. For simplicity, Table 3 provides an overview of the actual distances from each of the source words to the blends in our dataset. This metric does not however tell one which source word contributed more to the resulting blend.

Table 3: Levenshtein string edit distances between source words and blends

Distance between SW1 and blend	SW1	Blend	SW2	Distance between SW2 and blend	Closest match	Shorter source word
5	web	webinaar	seminaar	2	2	1
3	bewaar	bewarea	area	3	1	2
2	Zuma	Zupta	Gupta	1	2	1
5	Ramaphosa	Ramaforie	euforie	4	2	2
3	deur	deurnis	deernis	1	2	1
6	semi	semigrasie	emigrasie	1	2	1
6	dag	dagmerrie	nagmerrie	1	2	1
4	kak	kapoen	pampoen	2	2	1
9	dik	dikkelicious	delicious	3	2	1
3	Twitter	Twitterati	literati	3	1	1
3	Meng	Mengels	Engels	2	2	1
5	jags	jimpel	simpel	1	2	1
6	beter	betereinder	bittereinder	2	2	1
3	glitter	glitterati	literati	2	2	1
5	poes	poeslisie	polisie	2	2	1
1	sensitief	sensiteef	teef	5	1	2
7	Engels	Englikaans	Afrikaans	4	2	1
4	hoer	hoeranje	oranje	2	2	1
6	kuber	kuberkulose	tuberkulose	1	2	1
6	onder	onderweldig	oorweldig	3	2	1
7	poes	poeskantoor	poskantoor	1	2	1
7	Suip	Suip-Afrika	Suid-Afrika	1	2	1
6	miljoen	miljuisend	duisend	4	1	2
1	lankmoedig	lankmoerig	moerig	4	1	2
4	drama	dramedie	komedie	3	2	1
6	middagete	midbyt	ontbyt	3	2	2
1	mismoedig	mismoerig	moerig	3	1	2
5	pikkewyn	pikkewouter	kabouter	5	1	0
1	Saterdag	Saterdrag	drag	5	1	2
2	Ramaphosa	Ramaphala	Phala-Phala	6	1	1

From Table 3, we can calculate the average distance between the blends and source words one and two, respectively. On average, the distance between SW1 and the blends is 4.4, suggesting that SW2 (with an average distance of 2.7 from the blends) is on average closer to the blends.

In Table 4 *milj|uisend* is used as an example to illustrate another annotation of the blends that shows how its graphemes are sourced from the two separate words. The annotation method illustrated in Table 4 indicates more explicitly how each source word contributes to the

formation of the blend. In the table a '1' in the last row indicates that the grapheme is taken from SW1, a '2' indicates that it is taken from SW2, and a '3' would be indicative of graphemes that SW1 and SW2 share in the blend. In the case of *milj*|*uisend* there are now overlapping graphemes.

Table 4: Example of Afrikaans blend annotation using milj|uisend

Letter slot	1	2	3	4	5	6	7	8	9	10
Letters from SW1 not in blend					o	e	n			
Letters from SW1 in blend	m	i	l	j						
Letters from SW2 in blend					u	i	s	e	n	d
Letters from SW2 not in blend				d						
Annotation for letter blendtype	1	1	1	1	2	2	2	2	2	2

The annotation above does not tell one how many edits (or changes) are needed from SW1 and SW2 to arrive at the blend. Calculating the minimum number of edits needed in a string, whether it be additions, subtractions, or substitutions, can be done automatically, based on the Levenshtein string-edit distance. Instead of doing this manually for all the blends in the dataset, we use the *stringdist* function of the *stringdist* package (Van der Loo, 2014) in *R* to calculate the distances (or number of edits) between source words and blends. Using this function in *R* provides a similar distance as the distance showed in Table 3. It is worth noting, however, that this function in some cases underestimates the distance. This happens when there are characters in the rest of the source word that also occur in the blend – as is the case with *miljoen* (the "en" also occurs in *milj*|*uisend*, but clearly does not come from *mijloen*, but from *duisend*). In this case the distance is not calculated as 6 (as was the case in Section 3.2), but rather as 4: The "o" is replaced by "u", the "i", "s", and "d" are added and this amounts to 4. In these cases, we have manually changed the distances to rather correspond to the calculation in Section 3.2.

In addition to calculating the distance between the source words and the blend, we can also determine the similarity between a source word and the blend, *i.e.* the contribution of each source word to the blend. The contribution of a word is calculated through the similarity score based on the Levenshtein distance between the two words. The similarity score is calculated as $1 - \frac{d_{Levenshtein}(SW,\ Blend)}{length\ (Blend)}$ and is expressed as a percentage of the number of characters in the blend. This can be done using the *levenshteinSim* function of the *RecordLinkage* package (Sariyar and Borg, 2022) in *R* or the *stringsim* function in the *stringdist* package (Van der Loo, 2014). However, in the same way in which the automated function underestimates the distance, these functions overestimate the similarity, because they count all the similar characters in the two words. In the case of *miljoen* and *milj*|*uisend* then, the "en" in *miljoen* also counts towards the similarity score, and therefore a score of 60% is obtained. We worked around this with a function that only compares the similar strings in the source words and blends. This means that as soon as a letter is introduced in the blend that is not in SW1, when reading from left to right, it signifies the end of the contribution for that source word. A similar

approach is then followed backwards for SW2. The results for all the blends are shown in Table 5 where we show how much each of the source words contribute to the blend.

Table 5: Contribution of source words to the blends

Contribution of SW1 to blend	SW1	Blend	SW2	Contribution of SW2 to blend	Source word that made the largest contribution	Shorter source word
38%	web	webinaar	seminaar	63%	2	1
57%	bewaar	bewarea	area	57%	0	2
40%	Zuma	Zupta	Gupta	80%	2	1
44%	Ramaphosa	Ramaforie	euforie	56%	2	2
57%	deur	deurnis	deernis	57%	0	1
40%	semi	semigrasie	emigrasie	90%	2	1
33%	dag	dagmerrie	nagmerrie	89%	2	1
33%	kak	kapoen	pampoen	67%	2	1
25%	dik	dikkelicious	delicious	67%	2	1
70%	Twitter	Twitterati	literati	60%	1	1
57%	Meng	Mengels	Engels	71%	2	1
17%	jags	jimpel	simpel	83%	2	1
45%	beter	betereinder	bittereinder	82%	2	1
70%	glitter	glitterati	literati	60%	1	1
44%	poes	poeslisie	polisie	56%	2	1
67%	sensitief	sensiteef	teef	44%	1	2
30%	Engels	Englikaans	Afrikaans	60%	2	1
50%	hoer	hoeranje	oranje	63%	2	1
45%	kuber	kuberkulose	tuberkulose	91%	2	1
45%	onder	onderweldig	oorweldig	64%	2	1
36%	poes	poeskantoor	poskantoor	73%	2	1
36%	Suip	Suip-Afrika	Suid-Afrika	64%	2	1
40%	miljoen	miljuisend	duisend	60%	2	2
70%	lankmoedig	lankmoerig	moerig	60%	1	2
50%	drama	dramedie	komedie	63%	2	1
50%	middagete	midbyt	ontbyt	50%	0	2
67%	mismoedig	mismoerig	moerig	67%	0	2
55%	pikkewyn	pikkewouter	kabouter	45%	1	0
67%	Saterdag	Saterdrag	drag	44%	1	2
67%	Ramaphosa	Ramaphala	Phala-Phala	44%	1	1

From Table 5, we can see that, for the 30 instances, SW1 makes the biggest contribution to the blends in 5 cases, whereas SW2 contributes most in 21 of the cases. In four cases, the two source words contribute equally to the blend. On average SW1 accounts for 53% of the structure of the blends, while SW2 on average accounts for 71% of the structure of the blends. This means that graphemes from SW2 are more common in the blends.

Considering the length of the source words, we can see that SW1 is the shortest word in most of the cases (21 out of the 30), SW2 is the shortest in 7 cases and in two cases (*milj*|*uisend* and *pikkew*|*outer*) the two source words are equal in length. Furthermore, it seems like in only four cases (*Rama*|*forie, milj*|*uisend, pikkew*|*outer* and *Rama*|*phala*) the shortest word was the word that contributed most to the blend. From the observational data therefore, it does not seem to be the case that the shortest word necessarily contributes the most to the blend. With more data, one would be able to also test this statistically.

Another question then arises as to how much of themselves each of the source words contribute to the blend. We show this with *milj*|*uisend* again*:* In the case of *milj*|*uisend* the word *miljoen* contributes $\frac{4}{7}$ of itself to the blend (the four characters *milj* out of the seven characters of *miljoen*), while *duisend* contributes $\frac{6}{7}$ (the six characters *uisend* out of the seven characters of *duisend*). As such, we can also determine how much of each source word is present in the blends. This is shown in Table 6.

Table 6: Percentage of each source word present in the blends

Percentage of SW1 present in blend	SW1	Blend	SW2	Percentage of SW2 present in blend	Source word with highest percentage present in the blend	Shorter source word
100%	web	webinaar	seminaar	63%	1	1
67%	bewaar	bewarea	area	100%	2	2
50%	Zuma	Zupta	Gupta	80%	2	1
44%	Ramaphosa	Ramaforie	euforie	71%	2	2
100%	deur	deurnis	deernis	57%	1	1
100%	semi	semigrasie	emigrasie	100%	0	1
100%	dag	dagmerrie	nagmerrie	89%	1	1
67%	kak	kapoen	pampoen	57%	1	1
100%	dik	dikkelicious	delicious	89%	1	1
100%	Twitter	Twitterati	literati	75%	1	1
100%	Meng	Mengels	Engels	83%	1	1
25%	jags	jimpel	simpel	83%	2	1
100%	beter	betereinder	bittereinder	75%	1	1
100%	glitter	glitterati	literati	75%	1	1
100%	poes	poeslisie	polisie	71%	1	1
67%	sensitief	sensiteef	teef	100%	2	2
50%	Engels	Englikaans	Afrikaans	67%	2	1
100%	hoer	hoeranje	oranje	83%	1	1
100%	kuber	kuberkulose	tuberkulose	91%	1	1
100%	onder	onderweldig	oorweldig	78%	1	1
100%	poes	poeskantoor	poskantoor	80%	1	1
100%	Suip	Suip-Afrika	Suid-Afrika	64%	1	1
57%	miljoen	miljuisend	duisend	86%	2	2

Percentage of SW1 present in blend	SW1	Blend	SW2	Percentage of SW2 present in blend	Source word with highest percentage present in the blend	Shorter source word
70%	lankmoedig	lankmoerig	moerig	100%	2	2
80%	drama	dramedie	komedie	71%	1	1
33%	middagete	midbyt	ontbyt	50%	2	2
67%	mismoedig	mismoerig	moerig	100%	2	2
75%	pikkewyn	pikkewouter	kabouter	63%	1	0
75%	Saterdag	Saterdrag	drag	100%	2	2
67%	Ramaphosa	Ramaphala	Phala-Phala	36%	1	1

In Table 6 we see the opposite of what we have seen in Table 5: In Table 6, SW1 contributes more of itself to the blend in 18 of the 30 cases, whereas SW2 contributes more of itself in 11 of the cases. In only one case (*semi*|*grasie)* both source words are represented completely in the blend. Considering the lengths of the words, it seems that in 24 cases, the shortest source word contributes more of itself to the blend. This makes sense however, since the shorter word has less to contribute, and we will not necessarily see this as a phenomenon that is exclusive to blending.

4 Conclusion and future work

The aim of this article has been to offer a basis for the description of lexical blends in Afrikaans. As part of laying the foundation for the discussion it is apparent from the literature that uncertainty remains where blends are concerned. After considering some of the literature it was paramount to source available blend constructions in Afrikaans. The dataset was analysed using Levenshtein distances to measure the distances between source words and blends. This approach was taken to evaluate whether it could serve as a mechanism to address the “descriptive problem” (Bauer 2012, 21) that blends pose.

In our article we set out to determine whether a hypothesis of Wulff & Gries also holds for Afrikaans blends, namely that the shorter source word contributes more to the blend than the longer source word. It is not possible to make broad generalisations due to the size of the dataset, but from our 30 constructions it was apparent that the shorter source words tend to be included without truncation in most blends (as described in Table 6). But in terms of contributions to the final blend, our data show that the longer source word is more strongly represented, as can be seen from Table 5.

Another hypothesis of Wulff & Gries, viz. that the blend maximises similarities between the source words, can also be tested in future work. The average distance between SW1 and blends is 4.4, while the average distance from SW2 to blends is 2.7 – as was seen from Table 3. More data (and, as Wulff and Gries showed, different types of data) would be needed to test this hypothesis. Attention could also be given to describing the meaning contained in Afrikaans blends in future studies.

References

Balteiro, Isabel. 2018. “Emerging hybrid Spanish–English blend structures: ‘Summergete con socketines’.” Lingua 205:1-14. https://doi.org/10.1016/j.lingua.2017.12.010.

Bauer, Laurie. 2006. “Compounds and minor word-formation types.” In *The Handbook of English Linguistics*, edited by Bas Aarts and April McMahon, 483-506. Oxford: Blackwell Publishing Ltd.

Bauer, Laurie. 2012. “Blends: Core and periphery.” In *Cross-disciplinary perspectives on lexical blending*, edited by Vincent Renner, François Maniez, and Pierre Arnaud, 11-22. Berlin/Boston: Walter de Gruyter.

Beliaeva, Natalia. 2016. “Blends at the intersection of addition and subtraction: Evidence from processing.” SKASE Journal of Theoretical Linguistics 13(2):23-45. http://www.skase.sk/Volumes/JTL32/pdf_doc/02.pdf

Beliaeva, Natalia. 2022. Is Play on Words Fair Play or Dirty Play? The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse. In *The grammar of hate: morphosyntactic features of hateful, aggressive and dehumanizing discourse*, edited by Natalia Knoblock,177-196. Cambridge: Cambridge University Press.

Cacchiani, Silvia. 2016. “On Italian lexical blends: Borrowings, hybridity, adaptations, and native word formations.” In *Crossing languages to play with words: Multidisciplinary perspectives*, edited by Sebastian Knospe, Alexander Onysko, and Maik Goth, 305-335. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110465600-015

Cannon, Garland. 1986. “Blends in English word-formation.” Linguistics, 24(4):725-753. https://doi.org/10.1515/ling.1986.24.4.725.

Carter, Ronald. 2004. Language and creativity: The art of common talk. London: Routledge.

Combrink, Johan, GH. Afrikaanse morfologie: capita exemplaria. Pretoria: Academica.

Connolly, Patrick. 2013. “The innovation and adoption of English lexical blends.” JournaLIPP 2:1-14. https://doi.org/10.5282/JOURNALIPP/68.

Gries, Stefan Th. 2004a. Isn’t that fantabulous? How similarity motivates intentional morphological blends in English. In *Language, culture, and mind*, edited by Michael Achard and Suzanne Kemmer, 415-428. Stanford, CA: CSLI.

Gries, Stefan Th. 2004b. “Shouldn’t it be breakfunch? A quantitative analysis of blend structure in English.” Linguistics 42(3):639-667. https://doi.org/10.1515/ling.2004.021.

Gries, Stefan Th. 2006. “Cognitive Determinants of Subtractive Word Formation: A Corpus-based Perspective.” Cognitive Linguistics 17(4):535-558. https://doi.org/10.1515/COG.2006.017.

Gries, Stefan Th. 2012. Quantitative corpus data on blend formation: Psycho-and cognitive-linguistic perspectives. In *Cross-disciplinary perspectives on lexical blending*, edited by Vincent Renner, François Maniez, and Pierre J.L. Arnaud,145-167. Berlin / New York: Mouton de Gruyter.

Jurado, Alejandro Barrena. "A study on the ‘wordgasm’: the nature of blends’ splinters." Lexis. Journal in English Lexicology 14 (2019). https://doi.org/10.4000/lexis.3916.

Jurafsky, Daniel & Martin, James H. 2008. Speech and language processing. New Jersey: Prentice-Hall.

Kelly, Michael H. 1998. “To “brunch” or to “brench”: Some aspects of blend structure.” Linguistics 36(3):579–590. https://doi.org/10.1515/ling.1998.36.3.579.

Kemmer, Suzanne. 2003. Schemas and lexical blends. In *Motivations in language: studies in honor of Günter Radden*, edited by Hubert Cuyckens, Thomas Berg, René Dirven, and Klaus-Uwe Panther, 69-98. Amsterdam: John Benjamins.

Kjellander, Daniel. 2019. “Gold Punning: Studying Multistable Meaning Structures Using a Systematically Collected Set of Lexical Blends.” Lexis 14. https://doi.org/10.4000/lexis.3962

Kjellander, Daniel. 2022. “Ambiguity at work: Lexical blends in an American English web news context.” Umeå: Umeå University. (PhD).

Kubozono, Haruo. 1990. “Phonological constraints on blending in English as a case for phonology-morphology interface.” Yearbook of Morphology 3:1-20.
Lalić–Krstin, Gordana, Nadežda Silaški, and Tatjana Đurović. (2023). Meanings of -nomics in English: From Nixonomics to coronanomics: How -nomics has extended its original meaning to additional senses. English Today 39(2):141-148. doi:10.1017/S0266078422000013.
Levenshtein, Vladimir. 1966. “Binary codes capable of correcting deletions, insertions, and reversals.” Doklady Akademii Nauk SSSR 163(4):845-848.
Micheli, M Silvia. 2022. An extensive analysis of blending in Contemporary Italian. Lingua 273:103341. https://doi.org/10.1016/j.lingua.2022.103341
Mohsin, Ekhlas Ali. 2020. Blend formation tendencies, from English to Arabic: a comparative study. Newcastle: Newcastle University. (PhD).
Monakhov, Sergei. 2021. “Collective language creativity as a trade-off between priming and antipriming.” Plos one 16(11):p.e0259285. https://doi.org/10.1371/journal.pone.0259285
Renner, Vincent. 2015. "Lexical blending as wordplay." In *Wordplay and metalinguistic/metadiscursive reflection: authors, contexts, techniques, and meta-reflection*, edited by Angelika Zirker and Esme Winter-Froemel, 119-133. Berlin: Walter de Gruyter.
Renner, Vincent. 2019. “French and English lexical blends in contrast.” Language Contrast 19(1):27-47. https://doi.org/10.1075/lic.16020.ren.
Sariyar, Murat and Andreas Borg. 2022. RecordLinkage: Record Linkage Functions for Linking and Deduplicating Data Sets. Rnpackage version 0.4-12.4, https://CRAN.R-project.org/package=RecordLinkage.
Trollip, Benito and Trudie Strauss. 2023. Afrikaans lexical blends dataset. South African Centre for Digital Language Resources (SADiLaR), North-West University. https://repo.sadilar.org/handle/20.500.12185/668.
Van der Loo, Mark. 2014. “The stringdist package for approximate string matching.” The R Journal. 111-122. https://CRAN.R-project.org/package=stringdist.
Van Huyssteen, Gerhard Beukes. 2017. “Morfologie.” In *Kontemporêre Afrikaanse Taalkunde*, edited by Wannie Carstens and Nerina Bosman, 177-214. Pretoria: Van Schaik Uitgewers.
Van Huyssteen, Gerhard Beukes. 2020. “Subtraction.” Accessed 14 November 2022. https://taalportaal.org/taalportaal/topic/pid/topic-1571134395722850.
Villalva, Alina and Rafael Dias Minussi. "Description and analysis of a Portuguese blend corpus." Corpus 23 (2022). https://doi.org/10.4000/corpus.6436.
Virtuele Instituut vir Afrikaans (VivA). 2020a. “Korpusportaal: Omvattend.” Accessed 23 August 2023. http://korpus.viva-afrikaans.org/whitelab/search/simple.
Virtuele Instituut vir Afrikaans (VivA). 2020b. “Korpusportaal: Eksklusief.” Accessed 23 August 2023. http://korpus.viva-afrikaans.org/whitelab-versoek/search/simple.
Winters, Svitlana. 2017. Lexical Blending in Ukrainian: System or Sport? Calgary: University of Calgary. (PhD).
Wulff, Stefanie and Stefan Th. Gries. 2019. Improving on observational blends research: regression modeling in the study of experimentally-elicited blends. Lexis, 14. https://doi.org/10.4000/lexis.3625.