Compiling a candidate list of taboo constructions for an under-resourced language

Monique Rabé[1], Martin J Puttkammer[2], Gerhard B van Huyssteen[2]

[1] School of Languages, North-West University, Potchefstroom, South Africa
[2] Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa

E-mail: Monique.Rabe@nwu.ac.za, Martin.Puttkammer@nwu.ac.za, Gerhard.VanHuyssteen@nwu.ac.za

Abstract

Taboo-language resources remain scarce for under-resourced languages like Afrikaans – despite their clear relevance for natural language processing (NLP) and applications in artificial intelligence (AI). Although Afrikaans has a long-standing lexicographic tradition, it still lacks an open-access reusable lexical database for the taboo language. One of the most crucial steps in developing a constructional database for taboo language is to identify a candidate list of taboo constructions for potential lexicographic treatment. This paper outlines and tests a range of procedures to compile and refine such a list, with the goal of establishing a replicable methodology for similar work in other under-resourced languages. The methods draw on existing data of different types and corpora representing different registers. However, many entries are either false positives or ambiguous and require validation. Hence, we experiment with various semi-automated modelling techniques. These techniques include refining the candidate list through frequency analyses in corpora, expanding the list through partial corpus matching, and comparing the results against an attested, verified subset of taboo terms.

Keywords: Afrikaans; candidate list; lexical database; taboo language; under-resourced languages

1. Introduction

Although there is a large oeuvre of lexicographic products in or related to Afrikaans (Beyer & Louw, 2022), it still lacks a dictionary of taboo language. Also, only a handful of metalexicographic publications have dealt with Afrikaans taboo language – all of them dating from more than 25 years ago (see, e.g., Dekker, 1991; Feinauer, 1981; Harteveld & Van Niekerk, 1995; Van Huyssteen, 1998). This contrasts with an ever-growing body of (meta)lexicographic work that has been done for many well-resourced languages, like English, Dutch, and German (cf., Hughes, 1998; Sacher, 2012; Sakwa, 2012; Seemann et al., 2023; Van Huyssteen & Tiberius, 2023; WAON, 2013; Ziem et al., 2019). Nonetheless, despite the clear relevance of multifunctional taboo-language

resources for natural language processing (NLP) and applications in artificial intelligence (AI) (for example, see Ramos et al., 2024 for an overview of different approaches using or creating such resources), these kinds of resources remain scarce for languages other than English. Afrikaans, in particular, remains under-resourced: it has only a few authoritative dictionaries, a handful of small corpora, and a proprietary list of taboo words that is not openly available in the public domain.

Specialised taboo language resources – referred to in this article as constructional databases for taboo language (CDTLs) – are easily findable, accessible, interoperable, reusable digital resources enriched with annotations. These annotations typically include semantic (e.g., senses), orthographic (e.g., variants), morphological (e.g., derivations and compounds), syntactical (e.g., multi-word units, phrasal, and clausal constructions), paralinguistic (e.g., emojis), and pragmatic information (e.g., taboo ratings) relevant to each construction. The types of constructions that should be included span swearing, cursing, profanity, and obscenity; euphemisms, dysphemisms, and orthophemisms (e.g., names for sex acts or genitalia); and constructions used as abusive or harmful language (e.g., slurs, insults, or impolite expressions). Moreover, such constructions should not only cover lexical items (words and multiword units), but also so-called constructional idioms (i.e., constructions with "open" slots, like [WH the X] for *what the fuck*, *where the hell*, *how the devil*, etc.).

Following Van Huyssteen and Tiberius (2023: 66), reporting on a similar project to compile a lexical database for Dutch called *TaboeLex*, one of the first steps in populating the database – after deciding on its general design – is to compile a combined candidate list of taboo constructions that should be considered for lexicographic treatment. Against this background, this paper will report on the steps taken to empirically compile a usage-based, authoritative candidate list of Afrikaans taboo constructions to be used in a CDTL for Afrikaans, with the broader aim of developing a replicable methodology that could be applied to similar efforts in other smaller, under-resourced languages.

One of the biggest challenges in compiling a candidate list lies in the fact that perceptions of tabooness vary significantly between individuals, demographic groups, and pragmatic contexts. To illustrate the complexity of the task, we asked six experienced linguists, with broadly similar demographic and professional backgrounds, to annotate a random sample of 199 potentially taboo words as "alwaysTaboo", "oftenTaboo", "sometimesTaboo", "rarelyTaboo", or "neverTaboo".[1] No additional context – such as the etymology, definitions, or example sentences – was provided. Fleiss' kappa revealed that there was only slight agreement between the annotators, $\kappa = .121$ (95% CI, .103 to .139), $p < .0005$. The individual kappa for "alwaysTaboo" was highest, $\kappa = .384$ (95% CI, .348 to .420), $p < .0005$, indicating fair agreement between the annotators on this category. However, for the other four categories only

[1] All six respondents are first-language Afrikaans speakers between the ages of 40 and 55 (three male and three female), currently working as linguists at tertiary institutions in South Africa.

slight agreement was observed.

This illustrates the problem of obtaining consistent tabooness ratings from a small number of human annotators/raters, likely due to the lack of usage context, subjective interpretation of categories, and differences in personal and social perceptions of tabooness. The problem is further compounded by the presence of ambiguous or multifunctional words (e.g., *kom* as a verb can mean either 'to come' or 'to cum, to ejaculate semen', but as a noun *kom* typically means only 'cum, semen'), as well as entries that are clearly irrelevant (e.g., the word *Afrikaans* is listed in Urban Dictionary as potentially offensive).

For more reliable ratings, one could increase the population sample or remove outliers and borderline cases (see Eiselen & Van Huyssteen, 2023). However, the time and money it would cost to annotate a list of, say, 3,000 potentially taboo words, would make such an annotation endeavour practically and financially unfeasible. Furthermore, the kind of large-scale crowdsourcing approach employed by Wiegand et al. (2018) is also not viable for Afrikaans, given the limited availability of qualified Afrikaans-speaking annotators on platforms like Prolific Academic or Mechanical Turk.

To address these challenges, we experiment with alternative, semi-automated modelling techniques and verification procedures to generate a verified, unambiguous candidate list. In Section 2, we outline the approach and sources used to compile the initial candidate list of taboo constructions for Afrikaans. Section 3 explores methods for refining and expanding the candidate list using different corpora, while Section 4 evaluates the refined results by comparing it with a subset of verified taboo terms. Section 5 concludes the paper, outlining avenues for future work.

2. Step 1: Compiling a draft candidate list for an Afrikaans CDTL

According to Kiefer and Van Sterkenburg (2003: 353), a candidate list forms an integral part of the macrostructure of any lexicographic project and that the choice of lemmas to be included vary from language to language and should align with the general design, scope, and aim of the, in our case, lexical database. In our list, different types of headwords are included, namely:

- reduction forms (e.g., *PK* for *poes+klap* lit. pussy+slap > 'bitch slap', or *HKGK* for *hier kom groot kak* lit. here comes big shit > 'we're in big trouble');
- subwords, including infixes (e.g., *·fokken·* in *on·fokken·moontlik* 'im-fucking-possible'), prefixoids (e.g., *poes÷* in *poes÷groot* lit. pussy÷big > 'very big'), and suffixoids (e.g., *÷kop* in *pampoen÷kop* 'pumpkin head');

- words, including simplexes (e.g., *piel* lit. cock/dick > 'penis'), and complexes (e.g., *piel+kop* lit. penis+head > 'glans');
- multi-word units (e.g., *roer jou gat* lit. move your arse > 'hurry up'); and
- constructional idioms (e.g., *om [iemand] vir 'n gat te vat* lit. to [someone] for an arse PTCL.INF take > 'to take [someone] for a ride').

For Afrikaans, where no dedicated taboo-language dictionary exists, our candidate list was compiled mainly based on general-purpose, non-taboo dictionaries. This differs from the approach by Wiegand et al. (2018), who used resources such as curated wordlists annotated for polar intensity and sentiment orientation (including many abusive terms) that often don't exist for under-resourced languages. Nonetheless, our approach reflects a recognised – although often overlooked – lexicographical practice noted by Lauder (2010: 222): to use existing dictionaries as sources for differently purposed lexicographic resources.

Consequently, we used various labels that might be assigned to taboo constructions in Afrikaans dictionaries as seed labels to extract 312 lemmas from the *Handwoordeboek van die Afrikaanse Taal* (HAT, 2015), and 1,958 terms from the *Woordeboek van die Afrikaanse Taal* (WAT, 2025). These labels include, for example, *plat* ('coarse'), *vulgêr* ('vulgar'), *neerhalend* ('derogatory'), *kragtaal* ('strong language'), *vloek* ('curse'), *rassisties* ('racist'), *seksisties* ('sexist'), *kwetsend* ('offensive'), and *sleng* ('slang').[2]

However, as Lauder (2010: 223) cautions, "[a] dictionary which uses another as its main source of data will represent a limited and probably idiosyncratic view of the lexicon of the language," since tradition cannot be assumed to reflect actual usage. Consequently, we further supplemented the list with 391 lemmas from the popular Afrikaans satirical blog *WatKykJy*, 219 lemmas from the *Etimologiewoordeboek van Afrikaans* (EWA, 2003), 129 entries parsed from Afrikaans entries in *Urban Dictionary*, and an additional 73 lemmas from other resources (i.e., personal observations). This resulted in a first version of a rather large candidate list consisting of 3,082 candidates. Since various candidates appear in more than one of these sources, a clean-up and deduplication process was performed, resulting in a refined list consisting of 2,701 candidates (see Table 1 below).

During this clean-up process, we also decided to exclude multi-word units and constructional idioms from the candidate list at first, since their syntactic variability and context dependence make them harder to systematically extract and classify using the same methods applied to single-word headwords and subwords (see, for e.g.,

[2] The HAT, regarded as the most reputable concise monolingual dictionary for Afrikaans since its inception in 1965, contains about 70,000 lemmas. By contrast, the WAT is a far larger multi-volume descriptive dictionary, under development since 1926, with its first volume published in 1951. At present, it includes around 250,000 lemmas but remains incomplete, covering entries only from A to U at the time of our experiment.

Bergenholtz & Gouws, 2008; Fellbaum, 2015; Gouws, 2003; Louw, 2006; Tiberius & Colman, 2023 for the various approaches adopted for the treatment of these construction types in lexicographic resources).

Source	Lemmas extracted
Woordeboek van die Afrikaanse Taal (WAT)	1,958
WatKykJy	391
Handwoordeboek van die Afrikaanse Taal (HAT)	312
Etimologiewoordeboek van Afrikaans (EWA)	219
Urban Dictionary	129
Personal observations	73
Total before clean-up	**3,082**
Minus duplicates and multi-word units	−381
TOTAL: Draft candidate list	**2,701**

Table 1: Lemmas extracted from different sources

For instance, the form *[iemand] sit die pot mis* (lit. [someone] sits the pot miss > '[someone] misses the point completely') is included in the candidate list exactly as it was extracted from a source, but it will almost never appear in corpora in this fixed syntactic form: *iemand* ('someone') might be substituted with a proper noun or other pronoun (e.g., ***hy*** *sit die pot mis* '**he** misses the point completely'), or an adverb could be inserted (e.g., *iemand sit die pot* ***heeltemal*** *mis* 'someone **completely** misses the point'). Likewise, a multi-word unit like *de fok* ('the fuck') appears as part of a range of constructional idioms, such as *wat/hoe/hoekom de fok* ('what/how/why the fuck'). In these cases, even the definite article can vary – for instance, *wat/hoe/hoekom* ***die*** *fok* ('what/how/why the fuck') – and the lexical item *fok* ('fuck') itself actually functions as an open slot in this WHX construction, which can be filled with other material to form variants like *wat de/die* ***hel*** ('what the **hell**') (see Van Huyssteen et al. in press).

3. Step 2: Refinement and expansion using corpus matching

After compiling our draft candidate list in Section 2 above, we draw on corpus data from various registers to refine and expand this list: we use frequency evidence to narrow down candidates (Section 3.2) and partial matching techniques to identify further lemmas worth including (Section 3.3), based on actual usage.

3.1 Corpora used

Chyba: zdroj odkazu nenalezen below provides an overview of the different corpora that were used in the study to retrieve the frequencies of the 2,701 lemmas in the draft candidate list, indicating their size (n), version, degree of editing, and whether they are deemed to contain potentially taboo or mostly only non-taboo lexical content. In total, the combined corpora amount to just over 243 million words, with roughly 53 million words from corpora identified as containing taboo language and about 190 million words from non-taboo corpora.

The largest single corpus included is the Language Commission Corpus (Taalkommissiekorpus) 1.2, which consisted of 45,527,164 words at the time of our lookups. This corpus covers a wide selection of genres, encompassing fiction (prose works like novels and short stories) and non-fiction. The non-fiction category is further divided into academic writings, such as theses, dissertations, academic journal articles, and study guides, in addition to non-academic works like newspaper and magazine articles and non-fiction books. For the lookups, the possibly taboo content from this corpus – i.e., the words from the fiction genre (5,816,225 words) – and non-taboo content – i.e., the words from the non-fiction genre (39,710,939 words) – were treated as two separate corpora.

Similarly, the NWU Commentary Corpus (Kommentaarkorpus) 2.2, which consists of informal, unedited Afrikaans from a popular online newspaper's social media platforms, was analysed by dividing potentially taboo and non-taboo content into separate corpora. The comments that were flagged as being potentially harmful and subsequently removed from this newspaper's social media platforms by a content moderator, as part of the content moderation process, were used as the possibly taboo content (11,078,014 words), while the comments that remained formed the non-taboo part of the corpus (36,945,687 words).

Corpus name	***n***	**Version**	**Degree of editing**	**Lexical content**
NWU/LAPA Corpus	19,985,287	1.6	Edited	Taboo
NWU Commentary Corpus (Kommentaarkorpus)	11,078,014	2.2	Unedited	Taboo

PUK/Protea Boekhuis Corpus	10,482,340	2.4	Edited	Taboo
Language Commission Corpus (Taalkommissiekorpus)	5,816,225	1.2	Edited	Taboo
NWU/ATKV Tienertoneel Corpus	4,040,567	1.4	Semi-edited	Taboo
WatKykJy Corpus	1,789,376	2.2	Unedited	Taboo
SUBTOTAL: Taboo	53,191,809			
NWU/Maroela Media Corpus	45,485,069	2.2	Edited	Non-taboo
Language Commission Corpus (Taalkommissiekorpus)	39,710,939	1.2	Edited	Non-taboo
NWU Commentary Corpus (Kommentaarkorpus)	36,945,687	2.2	Unedited	Non-taboo
RSG News Corpus	36,063,150	2.9	Edited	Non-taboo
Afrikaans Wikipedia Corpus	28,105,289	1.7	Semi-edited	Non-taboo
NWU/ATKV Taalgenoot Corpus	2,568,984	1.2	Edited	Non-taboo
NCHLT Corpus	1,521,965	1.2	Edited	Non-taboo
SUBTOTAL: Non-taboo	**190,401,083**			
GRAND TOTAL	**243,592,892**			

Table 2: Taboo and non-taboo Afrikaans corpora

3.2 Method 1: Refinement

Based on the frequency lookups in the corpora described above, we classify candidate words into five data-driven categories, based on their observed frequencies in the taboo and non-taboo corpora. Chyba: zdroj odkazu nenalezen below shows the number of words falling into each of these categories, with each category briefly described after the table.

Category	**Description**	***n***
alwaysTaboo	Words that appear only in taboo corpora (frequency =	280 (10.4%)

	0 in non-taboo corpora)	
oftenTaboo	Words that appear in both corpora but with a higher frequency in taboo corpora	558 (20.6%)
sometimesTaboo	Words that appear in both corpora with the same frequency	78 (2.9%)
rarelyTaboo	Words that appear in both corpora but with a higher frequency in non-taboo corpora	494 (18.3%)
neverTaboo	Words that appear only in non-taboo corpora (frequency = 0 in taboo corpora)	116 (4.3%)
Not attested	Words sourced from dictionaries or other sources that have 0 frequency in both taboo and non-taboo corpora	1,175 (43.5%)
	TOTAL	**2,701**

Table 3: Distribution of initial candidate words by taboo status based on frequency patterns across taboo and non-taboo corpora

- Words categorised as "**alwaysTaboo**" are those that appear only in taboo corpora (e.g., *rond+fok* lit. around+fuck > 'fuck around').

- "**oftenTaboo**" words occur in both taboo and non-taboo corpora, but with a clearly higher observed frequency in taboo corpora (e.g., *fokken* 'fucking').

- "**sometimesTaboo**" words appear in both corpora with the same observed frequency. Examples include *kloot*, which has various meanings in Afrikaans – such as 'round object,' 'tipping truck,' and 'testicle' – and *woestersous*, a popular misspelling of *worcestersous*, a type of sauce popular in South Africa that can also refer to lubrication in a taboo context.

- "**rarelyTaboo**" words appear in both corpora but with a clearly higher observed frequency in non-taboo corpora. Examples are *Afrikaans*, which mostly refers to the language itself but is also defined by Urban Dictionary as "a language so bad that if you speak it, you get AIDS" (sic), and *jong* 'young,' which is mostly used literally, though it can also be used derogatorily to refer to a brown or black man.

- "**notTaboo**" words appear solely in non-taboo corpora (e.g., *agurkie* 'gherkin', mostly used in the literal meaning referring to a small cucumber, although it is sometimes used peripherally as a euphemism for the clitoris).

- Finally, there is a residual category for words that do not appear in any of the corpora. Although these words were sourced from dictionaries and other reference works, they do not occur in actual usage in the corpora we consulted. A possible explanation for the relatively high number of unattested items is that the majority were sourced from the *Woordeboek van die Afrikaanse Taal* (WAT), a near-century-old descriptive dictionary that aims to be comprehensive in its inclusion of words from all varieties of Afrikaans. A significant portion of these entries reflect historical, dialectal, or highly specialised use, with some being evidently outdated. In addition, the WAT has historically relied on fit-for-purpose corpora specifically compiled for its editorial process, which we do not have access to. Another contributing factor may be the nature of the corpora themselves: since taboo words feature most often in spoken language, they are often under-represented in corpora built on written texts, regardless of genre. As such, there is no direct correlation between the corpora we used and those initially employed to justify the inclusion of these entries.

We refined our candidate list by retaining only the words in the "alwaysTaboo" (280 words) and "oftenTaboo" (558 words) categories, resulting in a new list with only 838 potential candidates (subsequently referred to as CL 1.0).

3.3 Method 2: Expansion

After we narrowed down our draft candidate list to 838 words in Section 3.2 above, we aimed to identify which taboo words were most strongly associated with taboo corpora in terms of actual usage. To do this, we calculated the odds ratio (OR) for each word in CL 1.0, a statistical measure based on the observed frequencies showing how much more likely a word is to occur in taboo corpora than in non-taboo corpora. From this analysis, we selected the top 50 words (see Table 4 below illustrating only the top 10) with the highest odds ratios.

Headword	**English gloss (literal > taboo sense)**	**Observed frequency (*n*): Taboo**	**Observed frequency (*n*): Non-taboo**	**Odds ratio**
piel · e	lit. cock · PL / dick · PL > 'penises'	322	2	576.31
fokken	lit. fucking > 'very; [interjection of intensification]'	7847	66	425.65

ge · fok	lit. PST · fuck > 'ruined / screwed'	76	1	272.04
be · fok	lit. be · fuck > 'awesome / mad about / obsessed'	557	8	249.23
nool	lit. idiot > 'fool / simpleton'	64	1	229.09
op+fok	lit. up+fuck > 'mess up / ruin'	213	4	190.61
fokol	lit. fuck all > 'nothing'	635	12	189.42
kont	lit. cunt > 'female genitalia'	222	5	158.93
poes	lit. pussy > 'female genitalia'	1029	24	153.47
piel	lit. cock / dick > 'penis'	325	8	145.42
tos	lit. toss > 'to masturbate; rubbish'	239	6	142.58

Table 4: Top 10 taboo candidate words ranked by odds ratio, with observed frequencies (n) in taboo and non-taboo corpora

These top 50 words were subsequently used to perform partial matching within the combined taboo corpus, allowing us to identify additional morphological and syntactic variants of the same lexical items. This included inflected forms – such as the separable verb *rond+**tos*** 'toss around', its past participle *rond+ge·**tos*** 'tossed around', and the nominalised form *rond+ge·**toss**·ery* 'fuckery' – all partial matches of *tos* 'toss'. It also included compounds with prefixoids, such as ***kak**÷snaaks* lit. shit÷funny > 'very funny', and ***poes**÷hard* lit. pussy÷hard > 'very hard'. This step served to expand the coverage of the candidate list by capturing non-identical but related forms. Although all 50 words were used as search terms, only 43 of them yielded partial matches.

As a result, and after removing duplicates (e.g., since a compound like *moeder+fokker* 'mother fucker' was yielded as a partial match for both *fok* 'fuck' and *fokker* 'fucker'), a total of 4,894 partial matches were identified. Figure 1 below provides a breakdown of the number of partials retrieved for each of the 50 candidate words.

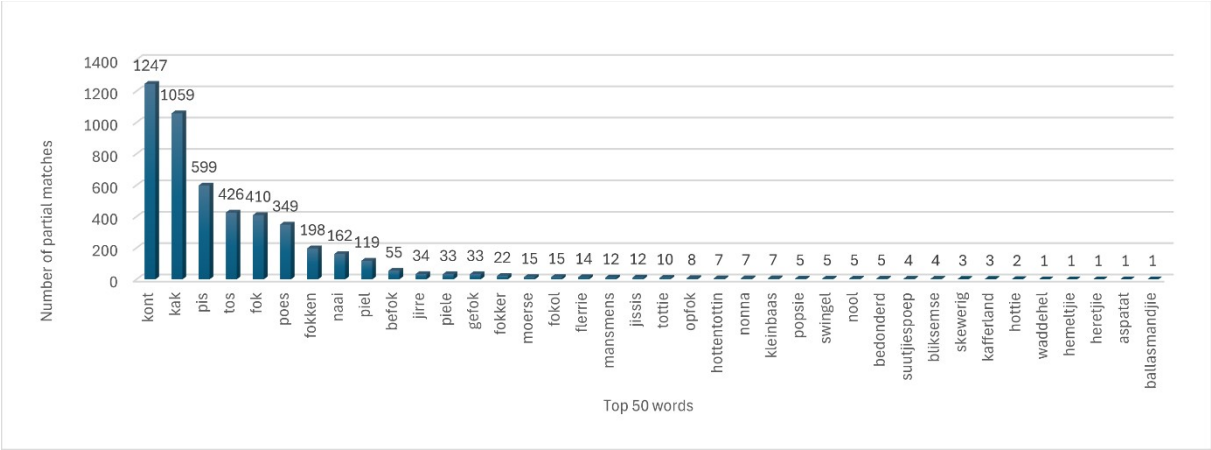

Figure 1: Number of partial matches found in taboo corpora for each of the top 50 candidate words

Hereafter, we repeated method 1 (see Section 3.2) to once again categorise the list of partial matches into the five data-driven categories. However, since the taboo corpora were used to identify the partial matches in the first place, no words were observed that would fall into the "neverTaboo" category. The distribution of the partial matches across the four observed categories is presented in Table 5 below.

Category	**Description**	***n***
alwaysTaboo	Words that appear only in taboo corpora (frequency = 0 in non-taboo corpora)	3,422 (69.9%)
oftenTaboo	Words that appear in both corpora but with a higher frequency in taboo corpora	301 (6.2%)
sometimesTaboo	Words that appear in both corpora with the same frequency	348 (7.1%)
rarelyTaboo	Words that appear in both corpora but with a higher frequency in non-taboo corpora	823 (16.8%)
	TOTAL	**4,894**

Table 5: Distribution of partial matches by taboo status based on frequency patterns across taboo and non-taboo corpora

Again, we refined the list to only the words that appear solely in the taboo corpora (i.e., "alwaysTaboo"; 3,422 words), and those that appear with a higher frequency in the taboo than the non-taboo corpora (i.e., "oftenTaboo"; 301 words). To further polish this list, we further filtered the "oftenTaboo" results by only keeping those where the

taboo frequency is at least double the non-taboo frequency (230 words). This resulted in a filtered candidate list of partial matches consisting of 3,652 words (subsequently referred to as CL 2.0).

After combining CL 1.0 and 2.0, and once again removing duplicates, this resulted in our final candidate list consisting of 4,422 words (hereafter referred to as CL 3.0; see Table 6 below).

Source version	Taboo status category	*n*
CL 1.0	alwaysTaboo (from draft candidate list)	280
	oftenTaboo (from draft candidate list)	558
SUBTOTAL: CL 1.0		**838**
CL 2.0	alwaysTaboo (partial matches of top 50 words from CL 1.0)	3,422
	oftenTaboo (partial matches of top 50 words from CL 1.0, taboo ≥ 2× non-taboo)	230
SUBTOTAL: CL 2.0		**3,652**
Combined total before deduplication		**4,490**
Minus duplicates across CL 1.0 and 2.0		−68
GRAND TOTAL: CL 3.0		**4,422**

Table 6: Compilation of final taboo candidate list by source and refinement stage

4. Step 3: Validation through comparison

As a last step to verify the validity of the entries in our final candidate list (i.e., CL 3.0), we compared it against a proprietary subset of manually checked taboo terms from the Centre for Text Technology (CTexT) of the North-West University. CTexT's offensive list is a proprietary, hand-curated list that originated during the manual review of the Afrikaans spelling checker's lexicon back in 2002. It was developed specifically with that application in mind – namely, to flag potentially offensive words and ensure that, while such words would never be suggested to users, their spelling

could still be verified if already present in a text. It has been subsequently enriched by incorporating entries from various dictionaries and partially matching potentially offensive terms found in corpora, all verified manually, over the last 20 years. At the time of our experiment, it included 2,968 entries. Upon reviewing the final candidate list against CTexT's list, only 601 words appeared in both.

While this might seem like a modest intersection, it is mainly due to the differing objectives and source materials: our candidate list encompasses various forms from informal, user-generated content on social media, which frequently includes spelling errors, neologisms, and complex morphological constructions. Nevertheless, the overlap of 601 items provides a robust, manually verified core set of prototypical Afrikaans taboo words, which can be used in the development of a CDTL for Afrikaans. For the remaining 3,821 candidate words that were not found in CTexT's list, we performed a qualitative inspection and found that most of them are indeed taboo and relevant in nature. This expansion is potentially valuable, as it adds newer or lesser attested forms that could enrich existing resources. We recommend that these words be submitted to a team of language experts for manual review before final inclusion.

5. Conclusion and future research

This article outlines a three-stage methodology for developing a candidate list of Afrikaans taboo constructions for potential inclusion in a constructional database for taboo language (CDTL); see Figure 2 below. Step 1 involved the compilation of a draft candidate list of 2,701 headwords from multiple lexicographic and other reference sources. In step 2, the candidate list compiled in step 1 was refined and expanded using two methods: frequency-based corpus filtering (method 1) was applied to identify a core list of 838 high-frequency taboo terms (CL 1.0), and partial matching in the taboo corpora (method 2) was used to expand the list by an additional 3,652 candidate forms (CL 2.0). Step 3 compared the final set of 4,422 taboo candidates (CL 3.0) to CTexT's proprietary list, revealing an overlap of 601 verified entries.

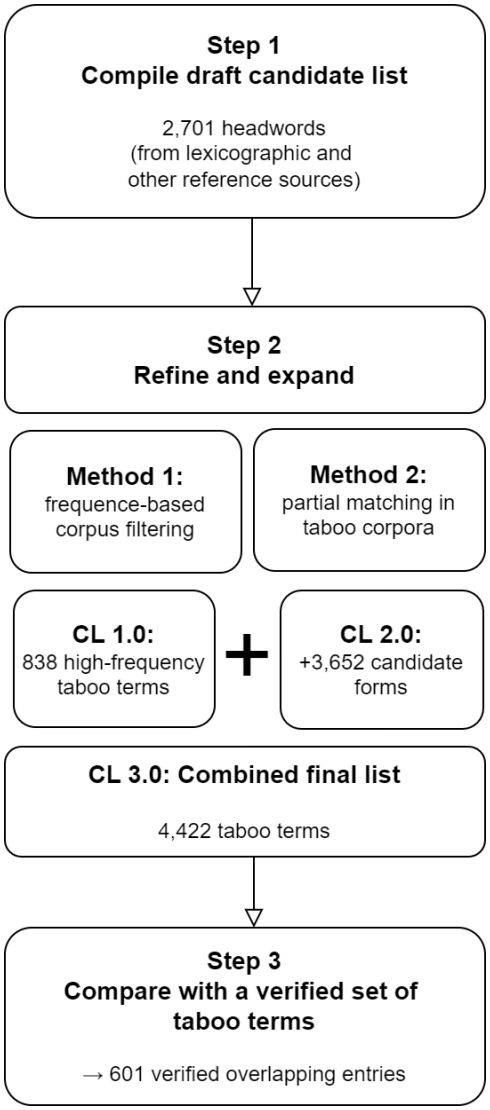

Figure 2: Three-stage methodology for compiling a candidate list of Afrikaans taboo constructions

Although the 601 resulting candidates offer a strong foundation for the development of an Afrikaans CDTL, it also underscores potential limitations of our current approach. More than 3,800 candidate words from our final list did not match CTexT's list. Yet, a manual inspection of these reveals that many are clearly offensive and appear to be used productively in online discourse (as detailed in Section 4). Their absence from the CTexT list may be attributed to two factors: (1) the high degree of morphological creativity and neologism formation observed in user-generated content, and (2) the informal and uncurated nature of the taboo corpora we chose.

These findings point to several avenues for future research. First, while our statistical filtering methods (based on frequency ratios and odds ratios) proved effective in narrowing down a candidate list, further experimentation with alternative semi-automated filtering techniques could help refine the boundary between genuinely taboo and merely informal or ambiguous terms. One related possibility would be to draw on

the microstructural seed labels used to extract candidate items from dictionaries, using them as the basis for a ranking system, according to the number and nature of the labels used. However, several potential challenges must be considered. For example, the *Woordeboek van die Afrikaanse Taal* (WAT) and *Handwoordeboek van die Afrikaanse Taal* (HAT) differ substantially in scale and coverage, which could influence attempts to rank lemmas: the amount, intensity, and application of labels may vary, and the consistency with which lexicographers have applied these labels over the past half-century is not guaranteed. In addition, lemmas sourced from other reference works often lack such labels entirely. Together, these factors suggest that while microstructural labels could serve as a useful supplement to corpus-based evidence, their usefulness as a stand-alone filtering tool remains uncertain and requires systematic evaluation.

Second, the corpora themselves should be critically reassessed. Since taboo words are far more common in spoken language, they are unlikely to be adequately represented in corpora based on written texts, regardless of genre. This may help explain why our draft candidate list of 2,701 words – compiled largely from dictionary data – produced only 838 potential candidates when checked against the corpora (see Section 3.2). Although we deliberately included informal genres such as online commentary to better reflect modern taboo language use, this choice may also have skewed the results toward more dynamic and rapidly changing expressions.

Third, although this study focused primarily on single-word forms, multi-word units and constructional units – which were deliberately excluded from detailed analysis here – constitute a vital and highly productive part of the taboo lexicon. As outlined in Section 2, these should be systematically analysed and verified as a separate category. Finally, we recommend that the remaining 3,821 candidate words, as well as all excluded multi-word units, be submitted to a team of language experts to verify their taboo status and suitability for inclusion.

Lastly, while the primary aim of this article was to develop a workable, semi-automated methodology for compiling a candidate list for an Afrikaans CDTL, we also emphasised a secondary aim: to create a replicable method for other under-resourced languages. While the scope of this article did not allow us to test the method on another small language, we believe that the first two steps can be applied to any language, provided that at least one monolingual dictionary, as well as a taboo and a non-taboo corpus, are available. Although our approach relied on a proprietary list of offensive terms to test the method, such a list is not a prerequisite for replication. We also acknowledge that some very small or relatively young languages may lack access to monolingual dictionaries or suitable corpora. However, such an absence of digital resources likely indicates that the language is still in the early stages of building a digital footprint and that a CDTL would currently have limited practical value.

6. Acknowledgements

This contribution appears under the auspices of a project (*What the Swearword! Multidiscipline research and scientific communication on cursing)*, whose overall ethics clearance was registered with the Ethics Committee for Language Matters (ECLM) of the North-West University on 21 May 2019 (registration number: NWU-00632-19-A7). More specifically, it forms part of a sub-project within this larger umbrella project, titled *Constructicographic aspect of the Afrikaans taboo constructicon*, which received separate ethical clearance from the ECLM on 13 May 2024 (registration number: NWU-01102-24-A7).

7. References

Bergenholtz, H. & Gouws, R.H. (2008). The access process in dictionaries for fixed expressions. *Lexicographica: International Annual for Lexicography/Revue Internationale de Lexicographie/Internationales Jahrbuch für Lexikographie*, 23, pp. 237–260.

Beyer, H.L. & Louw, P. A. (2022). Aspekte van vernuwing in die Afrikaanse leksikografie: Standaard- en pedagogiese woordeboeke as barometer [Aspects of innovation in Afrikaans lexicography: Standard and pedagogical dictionaries as a barometer]. *Lexikos,* 32(3), pp. 25–48. https://doi.org/https://doi.org/10.5788/32-3-1730

Dekker, L. (1991). Vloek, skel en vulgariteit: hantering van sosiolinguisties aanstootlike leksikale items [Swearing, cursing and vulgarity: dealing with sociolinguistically offensive lexical items]. *Lexikos,* 1(1), pp. 51–62. https://doi.org/https://doi.org/10.5788/1-1-1148.

Eiselen, R. & Van Huyssteen, G. B. (2023). A comparison of statistical tests for Likert-type data: the case of swearwords. *Journal of Open Humanities Data*, 9. https://doi.org/10.5334/johd.132

EWA: Etimologiewoordeboek van Afrikaans. (2003). Stellenbosch: Buro van die Woordeboek van die Afrikaanse Taal (WAT).

Feinauer, I. (1981). Die taalkundige gedrag van vloekwoorde in Afrikaans [MA dissertation, University of Stellenbosch]. Stellenbosch.

Fellbaum, C. (2015). The treatment of multi-word units in lexicography. In P. Durkin (ed.) *The Oxford handbook of lexicography.* Oxford University Press. pp. 411–424.

Gouws, R.H. (2003). Types of articles, their structure and different types of lemmata. In P. van Sterkenburg (ed.) *A practical guide to lexicography.* Amsterdam/Philadelphia: John Benjamins. pp. 34–43.

Harteveld, P. & Van Niekerk, A.E. (1995). Beleid vir die hantering van beledigende en sensitiewe leksikale items in die Woordeboek van die Afrikaanse Taal (WAT) [Policy for the treatment of insulting and sensitive lexical items in the Woordeboek van die Afrikaanse Taal (WAT)]. *Lexikos*, 5(5), pp. 232-248.

HAT: Handwoordeboek van die Afrikaanse Taal. (2015). Cape Town: Pearson.

Hughes, G. (1998). *Swearing: a social history of foul language, oaths and profanity in English*. London: Penguin Books.

Kiefer, F. & Van Sterkenburg, P. (2003). Design and production of monolingual dictionaries. In P. van Sterkenburg (ed.) *A practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins. pp. 350–365.

Lauder, A. F. (2010). Data for lexicography The central role of the corpus. Wacana: *Journal of the Humanities of Indonesia*, 12, pp. 219–242.

Louw, P. A. (2006). Inclusion strategies for multi-word units in monolingual dictionaries. *Lexikos*, 16(1), pp. 95–103.

Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C. & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1), pp. 204. https://doi.org/10.1007/s13278-024-01361-3

Sacher, J. (2012). *How to swear around the world*. San Francisco: Chronicle Books.

Sakwa, L. N. (2012). Problems of usage labelling in English lexicography. *Lexikos*, 21(1). https://doi.org/10.5788/21-1-47

Seemann, N., Lee, Y. S., Höllig, J. & Geierhos, M. (2023). Generalizability of abusive language detection models on homogeneous German datasets. *Datenbank-Spektrum*, 23(1), pp. 15-25. https://doi.org/10.1007/s13222-023-00438-1

Tiberius, C., & Colman, L. (2023). Lemmatisation of MWEs in Dutch resources [Poster]. 1st UniDive Workshop at Paris-Saclay, France.

Van Huyssteen, G.B. (1998). Die leksikografiese hantering van seksuele uitdrukkings in Afrikaans [The lexicographic handling of sexual expressions in Afrikaans]. *South African Journal of Linguistics*, 16(2), pp. 63–71. https://doi.org/https://doi.org/10.1080/10118063.1998.9724137

Van Huyssteen, G.B. & Tiberius, C. 2023. Towards a lexical database of Dutch taboo language. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (eds.) *Electronic Lexicography in the 21st century (eLex 2023): Invisible Lexicography*, eLex 2023. Brno, Czech Republic. pp. 53–74.

Van Huyssteen, G.B., Breed, A. & Pilon, S. in press. "'What the hell?!' vs. 'Wat de hel?!': Contrasting the intensifying WHX construction in English and Afrikaans." In D. van Olmen, M. Andersson, J. Culpeper & R. Giomi. (eds.) *The grammar of impoliteness: Trends in linguistics*. Berlin: De Gruyter Mouton.

WAON: Woordenboek van het Algemeen Onbeschaafd Nederlands. (2013). Houten/Antwerpen: Uitgeverij Unieboek | Het Spectrum bv.

WAT: Woordeboek van die Afrikaanse Taal. (2025). Stellenbosch: Buro van die Woordeboek van die Afrikaanse Taal (WAT).

Wiegand, M., Ruppenhofer, J., Schmidt, A. & Greenberg, C. (2018). Inducing a lexicon of abusive words – a feature-based approach. In M. Walker, H. Ji & A. Stent (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), NAAC HLT 2018. New Orleans, Louisiana. pp. 1046–1056.

Ziem, A., Flick, J. & Sandkühler, P. (2019). The German constructicon project: framework, methodology, resources. *Lexicographica*, 35(1), pp. 15–40.