http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

Semi-automating the Reading Programme for a Historical Dictionary Project

Tim van Niekerk, *Dictionary Unit for South African English, Rhodes University, Grahamstown, South Africa (dsae@ru.ac.za)*

Johannes Schäfer, *Department of Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany (johannes.schaefer@uni-hildesheim.de)*

and

Ulrich Heid, *Department of Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany (heidul@uni-hildesheim.de)*

Abstract: This paper describes the resources and software procedures used or developed in a major enabling step towards the revision of the scholarly reference work *A Dictionary of South African English on Historical Principles* (*DSAE*, Silva et al. 1996), namely the semi-automatic generation of a digitally-sourced lexical database on which new and updated dictionary entries will be based; as well as the addition, in parallel, of a new corpus of South African English (SAE) to the project. Drawing on online data sources and an extensive list of known SAE word forms, we have developed a software toolchain to gather, encode, annotate and collate textual sources, producing: (i) a 3.1-billion part-of-speech-annotated corpus of South African English; (ii) a lexical database of illustrative quotations for over 20,000 known SAE word forms, available for selection at the entry-revision stage; and (iii) a list of potential new variant spellings and headword inclusion candidates. These steps replace, where recent electronic sources are concerned, the mechanical aspects of quotation gathering, normally undertaken manually through a reading programme requiring years of teamwork to acquire sufficient coverage (cf. Hicks 2010).

Keywords: CORPORA, DICTIONARY WORKFLOWS, HISTORICAL LEXICOGRAPHY, LANGUAGE VARIETIES, LEXICAL DATABASES, READING PROGRAMMES, SOUTH AFRICAN ENGLISH

Opsomming: Die semi-outomatisering van die leesprogramme van 'n historiese woordeboekprojek. Hierdie artikel beskryf die hulpbronne en sagtewareprosedures wat gebruik word of ontwikkel is in 'n belangrike bemagtigingstap na die hersiening van die vakkundige naslaanwerk *A Dictionary of South African English on Historical Principles* (*DSAE*, Silva et al. 1996), naamlik die semi-outomatiese generering van 'n leksikale databasis van digitale bronne waarop nuwe en bygewerkte woordeboekinskrywings gebaseer sal wees; asook die gelyktydige toevoeging van 'n nuwe korpus van Suid-Afrikaanse Engels (SAE) tot die projek. Gebaseer op aanlyn data-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

bronne en 'n uitgebreide lys bekende SAE woordvorme, het ons 'n sagteware nutsketting ontwerp vir die versameling, enkodering, annotering en vergelyking van teksbronne, wat gelei het tot die skep van (i) 'n 3.1-biljoen woordsoortgeannoteerde korpus van Suid-Afrikaanse Engels; (ii) 'n leksikale databasis van illustratiewe aanhalings vir ongeveer 20,000 bekende SAE-woordvorme, wat by die hersieningsfase van die inskrywings beskikbaar is vir seleksie; en (iii) 'n lys van potensieel nuwe variante spellings en moontlikhede vir trefwoordseleksie. Wat onlangse elektroniese bronne betref, vervang hierdie stappe die meganiese aspekte van die versameling van aanhalings, wat gewoonlik met die hand met behulp van 'n leesprogram wat jare se spanwerk vereis om voldoende dekking te verkry, gedoen word (cf. Hicks 2010).

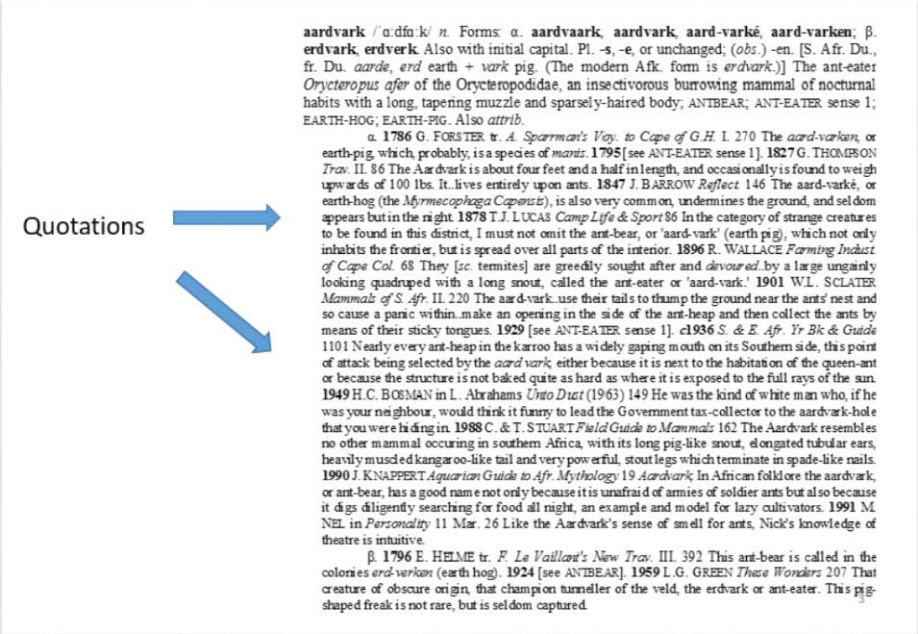
Sleutelwoorde: KORPORA, WOORDEBOEKWERKSVLOEI, HISTORIESE LEKSIKOGRAFIE, TAALVARIËTEITE, LEKSIKALE DATABASISSE, LEESPROGRAMME, SUID-AFRIKAANSE ENGELS

1. Role of quotations in the dictionary

A Dictionary of South African English on Historical Principles (*DSAE*, Silva et al. 1996) is a diachronic variety dictionary, first published as a single-volume print dictionary spanning about 800 pages and available as a pilot online edition at http://dsae.co.za since 2014. It closely resembles the *Oxford English Dictionary* (*OED*) in the design of its entries as well as its research processes, but focuses solely on South African English (SAE) from its origins in the late 17th Century onwards. The first edition of the *DSAE* was a long-term project involving three Editors-in-chief and 24 editorial staff and research assistants (excluding volunteer readers) over a period of 25 years. The result was a historical dictionary containing 4 600 main entries documenting about 17 500 word forms including headwords, plural forms, orthographic variants, compounds, phrases and derivatives. Of paramount importance are its evidential quotations (variously named contexts, citations or, informally among project staff, 'quots'). The quotations, drawn from monographs, periodicals, letters, manuscripts, ephemera and other sources are bibliographically referenced: while "in most other kinds of [monolingual] dictionary, attribution is rare … historical dictionaries generally provide information about the source and date of the quotation" (Atkins and Rundell 2008: 455). Much of the *DSAE*'s compilation process was therefore directed towards an ongoing reading programme. With the help of numerous volunteer readers, approximately 300,000 index card citations were collected as illustrative evidence for dictionary entries, their sense-divisions as they evolve through time, and nested lemmas. Of these about 45,000 quotations were included in the printed version of the dictionary, resulting in an average of 10 quotations per entry and producing a full running text of about 1,5 million words. The object was akin to the *OED*'s, following the principle that "[t]he dictionary should set forth the life history of each single word" (Willinsky 1994: 225), prompting an empirical methodology "based on the analysis of quotations from many textual sources" which "interprets the meanings of words in relation

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

to historical evidence of their past usage" (Brewer, 2007: 239). The *DSAE*'s inclusion policy was fundamentally quotation-driven: without 'quots' to present within the dictionary as attestations of usage, the compilers could not draft an entry or sense division. See Figure 1 for an example of the preponderance of citation evidence in a typical entry.



aardvark /ˈɑːdfɑːk/ *n.* Forms: α. **aardvaark**, **aardvark**, **aard-varké**, **aard-varken**; β. **erdvark**, **erdverk**. Also with initial capital. Pl. **-s**, **-e**, or unchanged; (*obs.*) **-en**. [S. Afr. Du., fr. Du. *aarde*, *erd* earth + *vark* pig. (The modern Afk. form is *erdvark*.)] The ant-eater *Orycteropus afer* of the Orycteropodidae, an insectivorous burrowing mammal of nocturnal habits with a long, tapering muzzle and sparsely-haired body; ANTBEAR; ANT-EATER sense 1; EARTH-HOG; EARTH-PIG. Also *attrib.*

α. **1786** G. FORSTER tr. *A. Sparrman's Voy. to Cape of G.H.* I. 270 The *aard-varken*, or earth-pig, which, probably, is a species of *mantis*. **1795** [see ANT-EATER sense 1]. **1827** G. THOMPSON *Trav.* II. 86 The Aardvark is about four feet and a half in length, and occasionally is found to weigh upwards of 100 lbs. It..lives entirely upon ants. **1847** J. BARROW *Reflect.* 146 The aard-varké, or earth-hog (the *Myrmecophaga Capensis*), is also very common, undermines the ground, and seldom appears but in the night. **1878** T.J. LUCAS *Camp Life & Sport* 86 In the category of strange creatures to be found in this district, I must not omit the ant-bear, or 'aard-vark' (earth pig), which not only inhabits the frontier, but is spread over all parts of the interior. **1896** R. WALLACE *Farming Indust. of Cape Col.* 68 They [*sc.* termites] are greedily sought after and *devoured*..by a large ungainly looking quadruped with a long snout, called the ant-eater or 'aard-vark.' **1901** W.L. SCLATER *Mammals of S. Afr.* II. 220 The aard-vark..use their tails to thump the ground near the ants' nest and so cause a panic within..make an opening in the side of the ant-heap and then collect the ants by means of their sticky tongues. **1929** [see ANT-EATER sense 1]. **c1936** *S. & E. Afr. Yr Bk & Guide* 1101 Nearly every ant-heap in the karroo has a widely gaping mouth on its Southern side, this point of attack being selected by the *aard'vark*, either because it is next to the habitation of the queen-ant or because the structure is not baked quite as hard as where it is exposed to the full rays of the sun. **1949** H.C. BOSMAN in L. Abrahams *Unto Dust* (1963) 149 He was the kind of white man who, if he was your neighbour, would think it funny to lead the Government tax-collector to the aardvark-hole that you were hiding in. **1988** C. & T. STUART *Field Guide to Mammals* 162 The Aardvark resembles no other mammal occuring in southern Africa, with its long pig-like snout, elongated tubular ears, heavily muscled kangaroo-like tail and very powerful, stout legs which terminate in spade-like nails. **1990** J. KNAPPERT *Aquarian Guide to Afr. Mythology* 19 *Aardvark*, In African folklore the aardvark, or ant-bear, has a good name not only because it is unafraid of armies of soldier ants but also because it digs diligently searching for food all night, an example and model for lazy cultivators. **1991** M. NEL in *Personality* 11 Mar. 26 Like the Aardvark's sense of smell for ants, Nick's knowledge of theatre is intuitive.

β. **1796** E. HELME tr. *F. Le Vaillant's New Trav.* III. 392 This ant-bear is called in the colonies *erd-verken* (earth hog). **1924** [see ANTBEAR]. **1959** L.G. GREEN *These Wonders* 207 That creature of obscure origin, that champion tunneller of the veld, the erdvark or ant-eater. This pig-shaped freak is not rare, but is seldom captured.

Figure 1: Example entry *aardvark* from the print edition of *A Dictionary of South African English on Historical Principles* (Silva et al. 1996) showing quotations from 1786 to 1991, separated by orthographic pattern (*aard-* vs *erd*-spelling)

2. The need for new quotations

Following the publication of the first edition of the print *DSAE*, after which the lexicography unit's focus shifted to synchronic dictionary projects not requiring citations, quotations continued being collected as part of an ongoing background reading programme, but on a much smaller scale. By 2004, index cards and various intermediate electronic wordprocessing formats had been replaced by an electronic lexical database allowing the capture, editing and annotation of quotation records in XML (eXtensible Markup Language) format for future use. Subsequently, as an early step towards revision of the historical dictionary, a data verification project was initiated to correct transcription errors in the 45,000 quotations used in the *DSAE*'s first edition, also stored in the new data-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

base. (For details of this painstaking and resource-intensive process, including additional information about source types and methodological considerations impacting on quotation collection, see Hicks 2010.) By 2017, the electronic database contained only about 9,000 new quotations, however. This small number is deceptive: it also contained a high proportion of new headword candidates (over 2,000). Nevertheless, quotation collection had suffered due to intervening dictionary projects, or the digitisation stages of the *DSAE*, having taken priority. A persistent limiting factor was that a large-scale reading programme requires staff to co-ordinate it, and capturing quotations manually, even with the help of assistants, is a highly labour-intensive task.

Nevertheless, just as the latest *OED* revision dedicated "a vast amount of well-directed energy" towards gathering new quotations (Brewer 2007: 241), so the *DSAE* revision requires increased data holdings of post-1995 citations. This applies not only to quotations for new SAE words not included in the first edition, but equally to new quotations for words already described in it, for several reasons: (1) recent citations for all entries should at least be reviewed, if not always necessarily included, to ensure that entries and their sense divisions are still up-to-date; (2) to support a focused review and potential redrafting of those entries labelled *rare*, *historical*, *obsolete*, *obsolescent* or *nonce* usage based on the limited evidence available at the time of compilation (bearing in mind that at that stage attestations could not be discovered via electronic retrieval systems); and (3) from the point of view of training a new team of lexicographers unfamiliar with the historical entry model and its complex styling policies, it may be preferable to begin by updating existing entries before drafting new ones.

3. Typical quotation-gathering stages

The selection of quotations to be reproduced in dictionary entries at the entry drafting stage will probably always require a human eye, and the current project does not attempt to replace editorial judgement. Most of the preceding stages of quotation gathering are, however, laborious and mechanical, namely:

(1) accessing and reading (scanning) texts for SAE words

(2) capturing quotations containing these words

(3) capturing date and source information

(4) verifying capture against sources to correct capture errors

(5) recording the relationships of word forms to parent dictionary entries (e.g. adding IDs, canonical forms and noting orthographic variants)

(6) anticipating as-yet unknown orthographic variants and repeating 1–5 above on discovery of new illustrative quotations.

In the toolchain described below all these stages are either wholly or partly automated, substantially increasing the dictionary project's quotation holdings,

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

now drawn from recent corpus sources, while dramatically reducing the labour involved. Additionally, we perform further computational steps to highlight potential new SAE terms within the corpus.

4. Input data sources

In 2009 it was reported that "there is no large corpus to represent South African English" (Pienaar and De Klerk 2009: 356) and, apart from proprietary, unfinished, or very small special-purpose corpora of under 1 million words, no others were available to suit the *DSAE*'s quotation-gathering requirements prior to the current collaboration. Additionally, in order to apply a toolchain to process the corpus, track relationships between word forms and extract quotations, a full dataset is required (rather than, for example, a web interface to a corpus allowing individual searches).

In building the SAE corpus, we draw on two sources of data, a newspaper corpus and a generic web corpus.

4.1 Newspaper Corpus

The newspaper corpus was created for quotation-gathering purposes from a suite of Perl programs[1] customised to crawl seven South African online newspapers between 2015 and 2017. After the resulting articles were fed through a parser to strip HTML markup, a further pre-processing step removed corpus noise such as boilerplate headers and footers unrelated to the article at hand, producing a corpus of about 6,5 million sentences or 100 million tokens. See Table 1 for specific counts across sources.

Although not as large as the web corpus described below, the newspaper corpus on its own provides a source of SAE quotations far exceeding the research data formerly available to the dictionary project. It also preserves contextual information in the original HTML versions of the articles such as embedded author details, when indicated, and typographic features such as italic font. (Italicisation of word forms is sometimes useful as an indicator that the author possibly regards a SAE word as a borrowed form, helping the lexicographer judge assimilation.) These features could not, however, be retained in the corpus-encoded and lexical database versions of the data since corpus encoding and other automatic processing steps required plaintext as input. The toolchain did, however, automatically add links to the source HTML as metadata, allowing the lexicographer to consult the original source if desired.

4.2 Web Corpus

The second dataset is a generic web corpus generated from .za domain sources by the NLP Group of the Computer Science Department at Leipzig University,

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

as part of its CURL (Crawling Under-Resourced Languages) project (see Goldhahn et al. 2012). The dataset was supplied already split into individual sentences and it does not distinguish between source types (e.g. newspapers vs blogs). Preprocessing steps such as the removal of HTML markup had already been performed on these data along with sentence segmentation. The order of the sentences is scrambled but each has an accession date and source URL, meeting the minimum requirements for an electronic citation. While a corpus of sentences may not satisfy the needs of linguists requiring more context for written utterances, this format is suitable for the historical dictionary project which requires only brief attestations. The Leipzig/CURL strategy of splitting articles into sentences was likewise adopted with the newspaper corpus described in 4.1 above. This was done for the sake of uniformity across the consolidated SAE corpus, and to simplify other automated processing steps including corpus encoding. The resulting corpus amounts to approximately 150 million sentences or 3 billion tokens, averaging 20 words per sentence. Table 1 below provides specific counts for 2011 through 2014.

Type	**Subcorpus source**	**Number of sentences**	**Number of tokens**
Newspapers:	*BusinessLIVE*	2,762,984	40,643,811
	Daily Maverick	320,273	7,331,568
	DispatchLive	117,752	2,481,062
	Independent Online	258,598	5,438,759
	SowetanLIVE	1,835,773	20,206,205
	The Citizen	368,182	7,881,382
	TimesLIVE	834,656	15,934,246
Subtotal (Newspapers)		*6,498,218*	*99,917,033*
Web (generic):	.za Domains 2011	3,870,783	74,114,784
	.za Domains 2012	2,784,879	53,248,634
	.za Domains 2013	50,191,936	1,031,432,748
	.za Domains 2014	91,728,781	1,823,257,689
Subtotal (Web)		*148,576,379*	*2,982,053,855*
SAE Corpus Totals		**161,572,815**	**3,081,970,888**

Table 1: Sentence and token counts for the Newspaper and Web subcorpora of the SAE corpus

5. Toolchain and its output

The overall toolchain is illustrated in Figure 2. Having described the input data sources, processing stages and the resulting tools and datasets are elaborated below.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

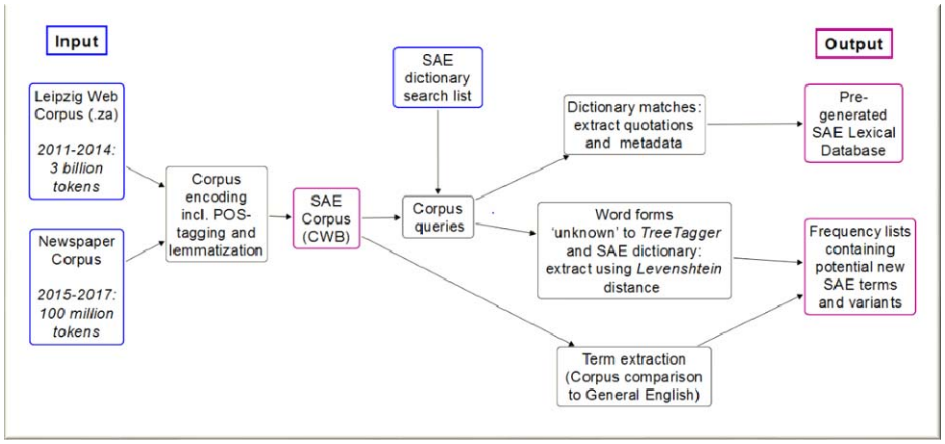


Figure 2: The full software toolchain showing the main inputs, outputs and processing stages

5.1 Annotated corpus and corpus query system

The toolchain introduces a corpus querying system to the *DSAE*'s set of research tools. Previously the project used a lexically and bibliographically annotated but comparatively miniscule lexical database of individually-captured quotations, or else Internet search engines. The lexical database, since it only contains quotations already captured, does not allow the discovery of new words or new senses of general English not yet recorded, and neither does this database nor general-purpose Internet searches allow queries according to linguistic attributes of word forms. Additionally, Internet sources are ephemeral, and continued access to quotations could previously only be assured by capturing them manually in the lexical database.

The SAE corpus solves these basic problems by providing a snapshot of the Internet across .za domains in a linguistically-annotated dataset that remains immutable even if the source web pages become inaccessible. In the preparatory stages, the dataset was part-of-speech-tagged (POS-tagged) and lemmatised using the TreeTagger[2], and loaded into the IMS Open Corpus Workbench (CWB)[3]. Past experiments at the lexicography unit showed that other concordancers like Wordsmith (Scott 2017) and AntConc (Anthony 2018) were not well-suited to the project's long-term needs. While these systems are user-friendly and helpful to many linguists, they "are designed to work with plain-text corpora … generally of rather small extent [and] they lack built-in support for complex annotation" (Evert and Hardie 2011: 2). Such annotation, along with search indexing, is required for sophisticated and efficient querying of very large datasets like the current one. The CWB and the query language implemented by its Corpus Query Processor (CQP) provide advanced search facilities such that *DSAE* editors can now disambiguate new usages from well-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

known ones. For example, the colloquial SAE verb *vrek* meaning 'die' is already described in the dictionary, with 11 quotations. The last one, dated 1990, reads: "English goes from bad to worse as ... Prince Charles rewrites Shakespeare — and Hamlet *vreks*" (*DSAE*, *vrek*, v.). The other use of *vrek* as an intensifier is not, however, recorded in the *DSAE*. Examples from the corpus show it paired with adjectives, e.g. *vrek dangerous* (extremely dangerous) or *vrek happy* (extremely happy), suggesting a dictionary update for this word. To find such cases previously, the editors would have had to resort to Internet phrase searches to find attestations, requiring that they imagine for themselves the possible alternative adjectives for *dangerous*, *happy* and so forth — a hit-and-miss methodology. Instead, the CQP tool now allows editors quickly to locate examples of *vrek* followed by any adjective.

At the same time, because the TreeTagger depends on an English lexicon that is relatively (and in some respects inevitably) unaware of SAE word forms, corpus annotations are sometimes incorrectly applied or simply lacking. For example, *vrek* is sometimes incorrectly tagged as a proper noun according to its predictive model, and because the tagger's lexicon does not contain an entry for this word or its inflections, the past participial form *vrekked* is never recognised as being associated with the lemma form. Searches for this word would sometimes therefore fail to return examples when queries are constrained strictly to part of speech or the canonical form. In these cases more relaxed query constraints over multiword contexts would, however, still typically produce significant numbers of usefully disambiguated results, and on such a large dataset, the advantages of this new research tool remain numerous.

5.2 Semi-automatically generated lexical database

5.2.1 General overview

The toolchain produced a second major result, namely the creation of a pregenerated lexical database compatible with its existing one. This resulted in a dramatic expansion of the project's electronically-encoded data holdings from about 9,000 quotations (captured manually between 2004 and 2017) to a gross count of about 147 million. This figure does include inordinately large sets of examples for common SAE words (e.g. 1315 quotations for *aardvark*) and SAE terms which overlap with general English (e.g. over 12,000 examples for *robot*, the SAE term for 'traffic light'). The quotations were extracted by matching a list of 21,718 SAE word forms against the entire corpus. This list, which was also used by the toolchain for other purposes, is described in more detail below (see 5.2.2 Input: SAE dictionary search list).

Table 2 shows frequencies of matches against the SAE search list by number of examples found. Despite occasional high-frequency matches attributable to terms which are also general English, most of the word forms searched for

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

are specific to the SAE lexicon and do not entail such overlap. Even if, for argument's sake, only 0.2% of the newly gathered quotations were considered, this would still approximate to the roughly 300,000 quotations gathered manually in the history of the project since it was established almost 50 years ago. When the corpus-derived quotations are reduced to 100 examples per word form, the total is about 540,000.

Number of quotations found per word	Number of words in search list
0	10,734
1-500	8,457
501-1000	526
> 1000	2,001
Total number of search terms	21,718

Table 2: Frequencies of matches against the SAE dictionary search list by number of quotations found

The lexical database with its new inclusions has a specific purpose for the dictionary: it differs from the CWB corpus in that it stores quotations already preformatted in the XML markup used to present quotations in the dictionary entries, with additional metadata for workflow and editing processes. The lexical database itself is designed to be interoperable with XML dictionary editing software and automatically renders quotations as HTML (Hypertext Markup Language) for published output. The XML output of the toolchain was therefore modelled in such a way that it was not only compatible with the existing database schema, but mirrored the basic hierarchical structure of the dictionary, facilitating further interoperability.

Figure 3 shows the lexical database's editing interface, with an automatically-generated record for *erdvark*, variant spelling of *aardvark*. The *e-* spelling was last recorded in the dictionary in 1959 (see Figure 1), making this 2014 instance a rare but valuable quotation. The record was generated by mapping the toolchain's XML output to the lexical database format via an XSL (Extensible Stylesheet Language) transformation. This data conversion process was fully automated, including the insertion of bibliographical information and annotations associating the *erdvark* quotation with various lexical attributes of its parent entry, as well as the more common spelling, ensuring easy retrieval during subsequent workflow stages. Roles such as *inputter*, *reader* and *annotation group* author have been performed by the toolchain and are therefore marked SYSTEM. The original source URL and the toolchain's corpus file are also included for reference.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

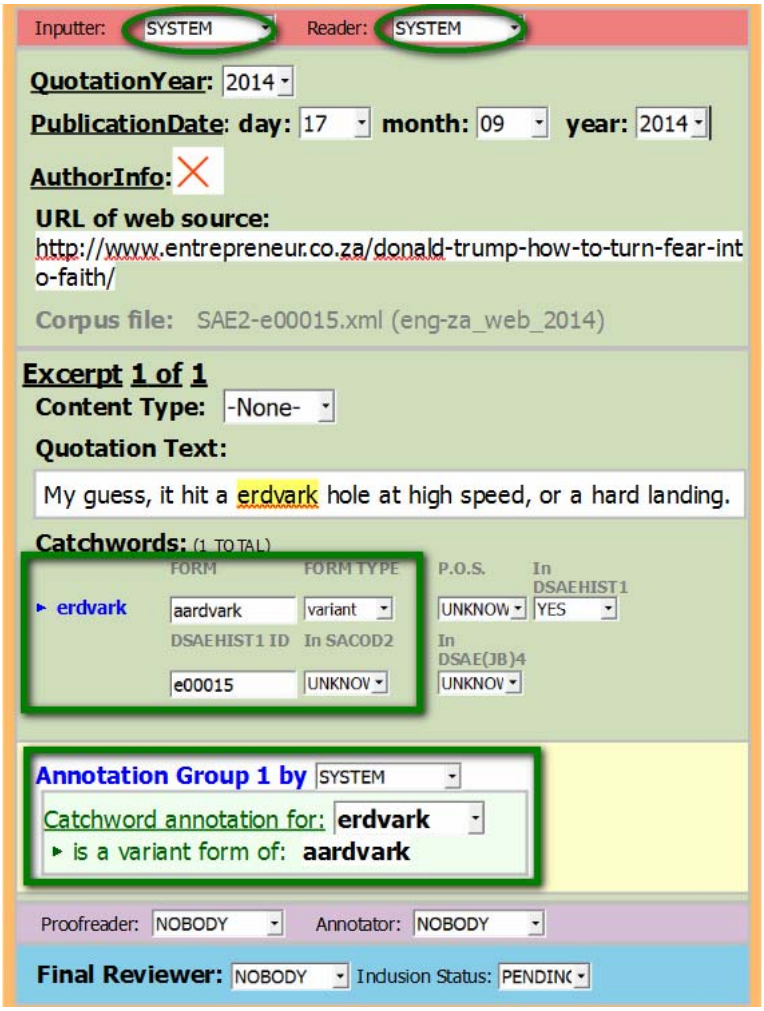


Figure 3: An automatically-generated lexical database entry for *erdvark*, variant spelling of *aardvark*, only requiring approval

The main remaining task of the human editor is to review the quotation and, if he or she considers it to be useful and reproducible in the dictionary, to update its *inclusion status* attribute from its default value PENDING to ACCEPTED. Acceptance requires that the quotation be checked for errors on the part of the author (but this proofreading stage no longer requires verification against the original source since capture was automated during corpus creation, removing the possibility of transcription errors). Further annotations may also be optionally added.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

Of course the editor should view all quotations with a critical eye, and in the current example the source URL happens to have been removed from the Internet since 2014. As with print-era ephemeral sources (leaflets, temporary signage and so forth) this does not mean the quotation cannot be cited. In this case the SAE corpus itself can ultimately be referenced: one of the advantages of the corpus is that it preserves ephemeral sources. It would also be desirable to note in the bibliographical metadata that this source comes from what used to be an industry news site, in this case for business entrepreneurs. This can be indicated manually in the lexical database interface via a SOURCE TYPE attribute not shown in the example. Such metadata could be added automatically in future by adding a subcomponent to the toolchain which queries a categorised list of the most frequently-cited domain names.

5.2.2 Input: SAE dictionary search list

The toolchain generates quotation records by matching word forms against the corpus using a simple string-matching process, drawing on a dictionary search list of 21,718 previously-documented SAE words. About 19,000 of these were drawn from the *DSAE*'s existing XML dataset which distinguishes between lemma types, namely *headwords* versus forms derived from these headwords (*variant spellings*, *plurals*, *compounds*, *derivatives* and other forms such as phrases). Included in this search list were hypothetical software-generated spellings for multi-word lexical items which could occur as orthographic variants due to hyphenation or spacing changes. For example, from the headword *mealie-meal* (901 corpus matches), *mealie meal* and *mealiemeal* were generated, producing 136 and 57 matches respectively. A further 2,393 new word forms were added to the *DSAE* list from existing post-2004 electronic holdings and categorised simply as *catchwords*. Although the catchword sub-list did not encode relationships between canonical and other forms, it brought valuable new or potentially-new words to the quotation-mining process. Table 3 shows the composition of the search list.

Type of word form	Number of items
Headword	6,057
Plural of headword	843
Variant spelling	7,444
Compound, derivative or other nested lemma form	4,981
Catchword (new words)	2,393
Total word forms	21,718

Table 3: Composition of SAE dictionary search list

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

Subsequent components of the toolchain centred on word forms based on their similarity to items in the SAE dictionary search list, or on the TreeTagger's failure to recognise them, in order to isolate lists of potential new SAE words or variants without dependence on pre-existing lexical knowledge.

5.3 Semi-automatic discovery of spelling variants and headword candidates

5.3.1 Analysis of new headword candidates unrecognised by the TreeTagger

Since the TreeTagger (see 5.1 above) relies on a general English lexicon for POS-tagging and lemmatisation, it assigned many words a default 'proper noun' POS value based on its probabilistic model, and an 'unknown' lemma value. As a precursor to further processing, a list of tokens with unknown lemmas tagged as proper nouns with a minimum frequency of 100 was analysed manually to assess the correctness of these POS-tags. This list amounted to 416 unique tokens. Review by an editor found 268 (64%) to have been correctly tagged as proper nouns. Further normalisation steps were also performed to improve overall results.

Normalisation and exclusion steps

Because list filtering was automated and frequency-based, two kinds of corpus preprocessing were undertaken before using the TreeTagger, to limit the number of irrelevant results. Firstly, numerous Unknowns took the form of a cardinal number followed by an alphabetical character as used in measures, e.g. *5 m*. Since the TreeTagger does not split these tokens and therefore cannot lemmatise them correctly, these were normalised in the corpus to add intervening whitespace, producing e.g.: *5 m.* Performing this step on an experimental subcorpus of about 2 million records, one per line, resulted in 2% of records being changed. Given that the Unknowns are a subset of otherwise correctly-lemmatised general English, this represented a significant reduction of corpus noise. A second preprocessing step used SAE proper noun exclusion lists to reduce the number of irrelevant Unknowns (typically proper nouns do not form part of the *DSAE*'s inclusion policy). A comprehensive list of South African Geographical Names[4] and a selective list of personal names, together totalling 47,987 single-word items and 1,027 multi-word proper names, were excluded. This list resulted in 625,787 unwanted corpus matches being filtered out.

5.3.2 Detection of new variants based on word similarity

The SAE dictionary search list included a list of documented variant spellings which were matched against quotations in the corpus. This left potential new

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

or previously undocumented variant forms which could be detected based on orthographic similarity between words in the corpus and words in the dictionary search list. Similarity was calculated using the Levenshtein distance algorithm, "a measure of the similarity between two strings … the source string (*s*) and the target string (*t*). The distance is the number of deletions, insertions, or substitutions required to transform *s* into *t*" (Gilleland n.d.).

This computationally-intensive process produced lists of word forms from the dictionary search list, each with a sub-list of similar words found in the corpus. These sub-lists were annotated and sorted first by Levenshtein distance measure, then by frequency in the corpus. See Table 4 for example data generated this way, showing potential variants of the SAE word *imphepho* (the name of a medicinal plant).

Words similar to *imphepho* (a medicinal plant), corpus frequency: 48		
Word form	**Frequency**	**Levenshtein distance**
impepho	77	1
imphepo	13	1
mphepho	4	1
imphephu	3	1
iphepho	3	1
mpepho	16	2
iphepha	15	2
iphupho	5	2
mphephu	5	2

Table 4: New variant candidates extracted from the corpus based on word similarity (likely candidates italicised)

Ordinarily, researching potential variant spellings is a painstaking process fraught with uncertainty. The lexicographer cannot easily anticipate all possible spelling permutations of borrowings from the several languages acting on English in the exceptionally multilingual context of South Africa. The pre-generated orthographic profile shown in Table 4 reduces labour, guesswork and subjectivity substantially, including the exclusion of imagined permutations which are not attested in the corpus, allowing quick evaluation of data in a single view. Being presented with multiple unfamiliar word forms may, at the same time, prompt unproductive corpus searches into what are found to be unrelated word forms not counting as valid SAE usage (e.g. code-switching). These may come as undesirable distractions increasing the burden of research. The lexicographer may, however, counter this with editorial judgment to compensate the Levenshtein algorithm's blindness to certain patterns, for example by

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

noting in this case that the most likely valid variant spellings are those which do not produce a vowel change.

The variant-tables also provide a quick indication of the relative currency of the headword's spelling form. For instance, the current example shows that one of the variants identified (*impepho*, 77 matches) occurs more frequently than the initial word form supplied as the search term (*imphepho,* 48 matches), suggesting that the former should be considered not as a variant but as the more likely headword candidate.

For shorter words, a Levenshtein maximum distance of 2 or 3 was found to be most productive in identifying new variant spellings. For longer words, typically compounds, a maximum distance of 4 or 5 was found to be useful. In the latter cases, variations due to spacing, hyphenation or lack thereof accounted for initial orthographic permutations, followed by further permutations possibly prompted by borrowing from a different language for part of the multiword item. For example, a search for *karretjie people* (SAE for a nomadic people who travel in animal-drawn carts or (Afrikaans) *karretjies*), with a maximum distance of 5, illustrated such alternations in language borrowing (*mense* is Afrikaans for 'people'):

(1) *karretjiepeople* (without space, Afrikaans + English)
(2) *karretjie-people* (hyphenated, Afrikaans + English)
(3) *karretjiemense* (direct borrowing of Afrikaans compound)
(4) *karretjie-mense* (ditto, hyphenated)
(5) *karretjies̲mense* (with Afrikaans-style plural marker)

The hyphenation in (4) above (frequency: 4) likely represents Anglicisation in SAE, since hyphenation of compounds is not typical in Afrikaans. Likewise the plural form in (5) (frequency: 4) would probably not have been anticipated by a native English-speaking lexicographer. The direct borrowing from Afrikaans in (3) was found to be most frequent (39 corpus matches).

5.3.3 Detection of new headword candidates based on word similarity

Because the word forms matched against had already been filtered and therefore tended to produce words specific to SAE, a side-effect of the variant-detection process was that unrelated but new headword forms were sometimes uncovered. For example, a search for variant spellings of *bogadi* (a traditional African wedding gift) returned a 'similar' word *moladi* (a system for rapid and inexpensive wall construction, designed and used in South Africa). Levenshtein distance was 2 with 120 corpus matches, suggesting headword candidate status. While such results were difficult to predict given that they were incidental to the actual purpose of this toolchain component, they presented an additional means of lexical acquisition and they are flagged for the editor's attention by high frequency values.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

5.3.4 Detection of headword candidates using term extraction

In the final stage of the toolchain, standard term extraction techniques (cf. Ahmad et al. 1994) were used to detect potential new headword candidates. The term extractor *TrEx_v5.9_sae*[5] was used. It compared the SAE corpus with the British National Corpus (BNC), since the latter would likely show lower frequencies for SAE tokens, and confined analysis to terms with a minimum corpus domain frequency of 10. This process compared the relative frequency of tokens in each corpus and extracted those more prominent in the SAE corpus. As output, the tool produced 15 lists, each representing a part-of-speech pattern, with the term, its frequency, an example quotation and other statistical data. Of these statistical rankings the most relevant was its *termhood* value. Termhood measures "try to identify candidate terms which are used [or] specialized in the domain as technical terms" (Schäfer 2015: 49), and the tool was used to test the hypothesis that this measure would also highlight SAE-specific words. Given the very large scale of the SAE corpus, and because general English terms had not yet been filtered out, the lists produced were unwieldy, generating a startling total of 2,615,854 candidate terms. Ranking these candidates using a combination of corpus frequency and termhood value, however, made new headword candidates accessible, and identified this toolchain component as a useful new mechanism in semi-automatic lexical acquisition. Some example new noun headword candidates discovered this way are:

(1) *braairoom* (an entertainment room used for indoor barbecues)
(2) *mokoro* (a type of canoe used in Botswana)
(3) *miombo (woodland)* (a Southern African vegetation type).

6. Re-orientation of reading programme prompted by semi-automation

The preceding discussion of the data resources newly available to the historical dictionary project, and the algorithms and output of the toolchain, together suggest a long-term review of the project's workflow and policies during its revision stage. Topics either not mentioned or only lightly touched on in this paper — being detailed and beyond its scope — are inclusion policy, criteria for SAE status, entry revision prioritisation, and the role of the lexicographer in assessing evidence. The new data and tools impact on all of these questions. For example, the print edition was compiled in a period when electronic sources were only starting to become accessible. Its inclusion policy for headwords and its high proportion of variant spellings may have been based on the reasonable assumption that more evidence existed than was available in the project's index card database. Now, faced with a massive influx of new data — albeit only for 2011–2017 sources — should it perpetuate the same inclusive approach when drafting new entries? Already over 2,300 new headword candidates had been identified prior to the semi-automation process, or roughly

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

half the number of headwords in the *DSAE*'s first edition (a 25-year project). Given the scale of new data, the number of new headword candidates may also increase dramatically, requiring prioritisation, probably also best done with further computational methods. In order to cope with large scale data the question becomes: which types of lexicographical tasks can be delegated to machines?

Responding to the analogous question "Will there be lexicographers in the year 3000?", posed in 1998 by Gregory Grefenstette, Michael Rundell observes that:

> From the standpoint of the editor and publisher, the shift to automation offers the prospect of producing a more diverse range of lexical resources without the enormous costs associated with conventional dictionary-making. It seems likely that, for the time being, there will be a central role for skilled lexicographers and editors. But their role is changing, from selecting and synthesising information, to 'editing' and validating choices already made by software. (Rundell 2012: 17)

The semi-automation of the *DSAE*'s reading programme prompts such a change in roles. The separation of automatic data collection processes from the closing steps requiring human judgement, as detailed above (5.2 Semi-automatically generated lexical database), are ample illustration of a shift from 'selecting and synthesising' quotations to 'editing' (if necessary) and 'validating' them.

The transition extends beyond validation, however, in that the toolchain's dependence on predetermined, categorised word forms signals a more general change in data collection strategy: project staff should orient their collection efforts towards lists. Whereas previously reading for neologisms or updated quotations typically involved (1) opportunistic reading across a wide variety of sources, (2) checking back against the database to avoid duplication, and (3) searching for further attestations if necessary to establish the currency of a potential new word candidate, the task is now simply to find a single example. On capturing a single instance of a potential new word, preferably with its canonical and inflected forms distinguished, the toolchain will be far faster and more comprehensive in sourcing new quotations. Likewise, if a word form is already known to the toolchain, it will already have extracted all possible quotations from the corpus, and to source new examples manually would duplicate effort. Likewise, orthographically-similar spelling forms could be detected by the toolchain as described above, again using the new list item as a starting point.

7. Conclusion

This paper has described the acquisition of new electronic data sources, their encoding as a very large, queryable part-of-speech-annotated SAE corpus, the subsequent dramatic expansion of a historical lexical database, and the provision of new headword and variant spelling candidates using a computational linguistic toolchain. The next steps for the dictionary project prior to revision

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

involve incorporating these results most usefully into future workflows. For example, the sometimes overwhelming numbers of quotations now provided for certain words could potentially be filtered for easier evaluation or disambiguated to reveal new, undocumented patterns of usage. Similarly, additional data sources could be added to the toolchain. These would, however, be improvements on a major development for a previously under-resourced dictionary project where data holdings were concerned; several highly-enabling steps towards semi-automatic 'reading' for lexicographic evidence have already been taken.

8. Acknowledgements

Thank you to Ms Heike Stadler, University of Hildesheim, Germany for generously adapting and maintaining her suite of Perl programs between 2015 and 2017 for the purposes of the collaborative research described in this article. Ultimately this produced the input data for the 100-million-token newspaper corpus described in 4.1 above. We also gratefully acknowledge the supply of data from the NLP Group, Department of Computer Science, University of Leipzig, Germany (see 4.2).

9. Endnotes

1. Developed by Ms Heike Stadler, University of Hildesheim, Germany for the collaboration described in this article. Please see 8. Acknowledgements.
2. The TreeTagger is made "freely available for research" at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
3. Available open-source under the Gnu General Public Licence at http://cwb.sourceforge.net/.
4. Released in 2011 by Statistics South Africa (see http://www.statssa.gov.za/?p=1341) and supplied to the project for research use.
5. Originally developed as part of the project described in Schäfer 2015 (see p. 87).

10. References

A Dictionary of South African English. [Online]. Available: http://dsae.co.za.

Ahmad, K., A. Davies, H. Fulford and M. Rogers. 1994. What is a Term? The Semi-automatic Extraction of Terms from Text. Snell-Hornby, M., F. Pöchhacker and K. Kaindl (Eds.). 1994. *Translation Studies: An Interdiscipline*: 267-278. Amsterdam: John Benjamins.

Anthony, L. 2018. *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University.

Atkins, B.T.S. and M. Rundell. 2008**.** *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Brewer, C. 2007. *Treasure-house of the Language: The Living* OED. New Haven/London: Yale University Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1468

Evert, S. and A. Hardie. 2011. Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference, ICC Birmingham, 20–22July 2011.* Birmingham: University of Birmingham. Accessed at http://eprints.lancs.ac.uk/62721/ [07/27/18].

Gilleland, M. n.d. *Levenshtein Distance, in Three Flavors.* Accessed at https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm [25/7/2018].

Goldhahn, D., T. Eckart and U. Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. Calzolari, N. et al. (Eds.). 2012. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 21–27, 2012* (LREC 2012): 759-765. Istanbul, Turkey: European Language Resources Association. Accessed at https://pdfs.semanticscholar.org/1b56/0f892432fb853d233c92f9294640bc91de3c.pdf.

Hicks, S. 2010. Firming up the Foundations: Reflections on Verifying the Quotations in a Historical Dictionary, with Reference to *A Dictionary of South African English on Historical Principles. Lexikos* 20: 248-271. Accessed at http://lexikos.journals.ac.za/pub/article/view/142 [03/12/2017].

Pienaar, L. and V. de Klerk. 2009. Towards a Corpus of South African English: Corralling the Subvarieties. *Lexikos* 19: 353-371. Accessed at http://lexikos.journals.ac.za/pub/article/view/444 [03/12/2017].

Rundell, M. 2012. The Road to Automated Lexicography: An Editor's Viewpoint. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 15-30. Oxford: Oxford University Press.

Schäfer, J. 2015. *Statistical and Parsing-based Approaches to the Extraction of Multi-word Terms from Texts: Implementation and Comparative Evaluation.* BSc Thesis. Stuttgart: Institute for Natural Language Processing (IMS), University of Stuttgart.

Scott, M. 2017. *WordSmith Tools*. Stroud: Lexical Analysis Software.

Silva, P., W. Dore, D. Mantzel, C. Muller and M. Wright (Eds). 1996. *A Dictionary of South African English on Historical Principles.* Cape Town: Oxford University Press.

Statistics South Africa. 2011. *South African Geographical Names Database*. See http://www.statssa.gov.za/?p=1341.

Willinsky, J. 1994. *Empire of Words: The Reign of the OED.* Princeton, N.J.: Princeton University Press. Accessed at https://books.google.co.za/books?id=UvCbv3ckRDkC&dq [07/27/18].