Enabling Selective Queries and Adapting Data Display in the Electronic Version of a Historical Dictionary

Tim van Niekerk, Heike Stadler, Ulrich Heid

Dictionary Unit for South African English, Grahamstown and
Department of Information Science and Natural Language Processing, University of Hildesheim
e-mail: t.vanniekerk@ru.ac.za, stadle@uni-hildesheim.de, heid@uni-hildesheim.de

Abstract

We report on an ongoing project aimed at making the scholarly reference work *A Dictionary of South African English on Historical Principles* (*DSAE*, Silva *et al*. 1996) accessible for selective online querying[1]. On the basis of the XML version of the current online edition of the dictionary, we perform two major steps towards this goal: (i) the design of interactive presentation devices for the dictionary, and (ii) the enhancement of the available lexicographic data categories. We take inspiration from the Function Theory of Lexicography (Tarp 2008), even though its main focus is on communicatively oriented dictionaries, while the *DSAE* primarily serves cognitive purposes. Since we are reporting on an ongoing project, we only present examples of the work.

Keywords: historical lexicography; electronic lexicography; print adaptation; user interfaces; XML

1 Introduction

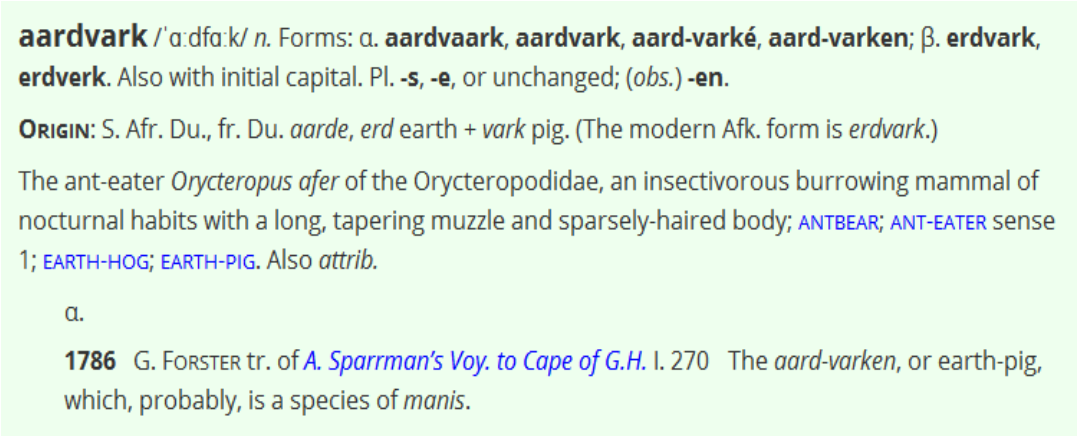
The *Dictionary of South African English on Historical Principles* (*DSAE*) is a variety dictionary with a strong emphasis on etymology and word history. It was first published in 1996 as an 850-page monovolume print dictionary with an entry design similar to that of the *Oxford English Dictionary* (*OED*), and has been out of print since 2004. In July 2014, a pilot online version was made available at http://dsae.co.za. Since the summer of 2015, the authors and their institutions have been adapting the online version to support selective querying and data display adapted for electronic platforms.

2 *DSAE Online*: Pilot Edition

The online edition of the *DSAE* provides access to 4,600 lemma entries, ca. 4,000 related word formation products and ca. 6,000 orthographic variants, a total of over 14,000 lexical items specific to the South African variety of English (SAE). Current access to the online *DSAE* is by incremental search with autocompletion prompts for lemmas and their orthographic variants, as well as via rudimentary full text search.

The dictionary has a rich microstructure: along with the lemma signs and their orthographic variants, inflected forms are also indicated where necessary, as well as grammatical categories, IPA transcriptions, detailed etymologies, pragmatic and diasystemic subvariety marks and domain labels. Meaning paraphrases and synonym indications are modelled on the treatment of polysemy found in the *OED*. Most importantly, the dictionary contains over 45,000 quotations, which are dated and bibliographically documented. Part of an entry from the current online version is reproduced in Figure 1.

[1] We gratefully acknowledge partial financial contributions to some of the work presented in this paper, from the Niedersächsische Staatskanzlei and MWK Niedersachsen (2015) and from the German Academic Exchange Service, DAAD (2015).



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**.

Origin: 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. of *A. Sparrman's Voy. to Cape of G.H.* I. 270 The *aard-varken*, or earth-pig, which, probably, is a species of *manis*.

Figure 1: Sample entry *aardvark* showing only the first of 16 illustrative quotations, from the pilot edition before improvement of online layout.

In terms of the Function Theory of Lexicography (Tarp 2008), the *DSAE* serves cognitive purposes, i.e. supplying information about words and things. User research shows that it is used by writers, editors and translators; by teachers and students at secondary schools and tertiary institutions; and by linguists and lexicographers. Another user group comprises laypersons interested in the history and culture of South Africa as well as those who discover the dictionary after querying Internet search engines on SAE lexical items.

3 Motivation and Objectives

The *DSAE* is the major source of information on South African English and an important source for English-based access to knowledge about South African culture, concepts and realia. Since its internal representation is already partly structured with lexicographic data categories, the objective of the work presented here is to complete as far as possible the contents-based markup, and simultaneously to produce innovative presentations of the data contained in the dictionary.

The procedures required in this context are as follows: (i) a semi-automatic explication of lexicographic data categories, as a prerequisite for (ii) adaptations to data display and (iii) design studies into the use of the dictionary by means of selective queries. In this article, we present examples of these steps.

In the remainder of this article, we give a reminder of selective querying (section 2), show examples of new data presentation devices (sections 3 and 4), and discuss examples of preparatory work on the explication of data categories in the *DSAE* (section 5). We conclude in section 6.

4 Towards Selective Querying and Data Display of Historical Dictionary Content

4.1 User Needs

Through ongoing work, and as an extension of the overall project of adapting the original print version for electronic platforms, we intend to broaden the range of possible uses of the *DSAE*. Currently, it

mainly serves (a) cognitive purposes, concerning both concepts (questions 1 and 2 below) and words (example 3), as well as (b) communicative purposes of text reception. Sample questions users may want answered by the dictionary are as follows:

(1) 'What does *aas-vogel* mean?' (historical spelling of *aasvoel*, the South African English term for the Cape vulture);
(2) 'Exactly what kind of plant is *melkbos*?' (a Dutch-derived vernacular name applied to various plants since the mid-19th-century in various English contexts);
(3) 'Is the word *baviaan* still used today; if not, what are its current synonyms?' (outdated; current words are *chacma (baboon)* or *bobbejaan*).

We intend to go further than this, by additionally enabling queries not related to a single lemma but to properties of lexical items (cf. examples 4 and 5). This presupposes a search across the dictionary based on any property or combination of properties of a lexical item, not only or not necessarily the lemma, e.g.:

(4) 'Which plant names in South African English derive from Khoikhoi?' (e.g. *buchu*, *dagga* or *kukumakranka*);
(5) 'When did Malay music terms enter South African English?' (e.g. *gom-gom* (1731), *ghomma* (1934), *klopse* (a synonym of *ghomma*; 1987) etc.).

4.2 Selective Querying and Data Display in Historical Dictionaries

In the 2014 online edition, users viewing the entries for the above-mentioned lemmas see the full text of the respective entry. As is customary with historical and scholarly dictionaries, richly documented entries of the *DSAE* may be rather lengthy and structured in a complex way: for example, the entry s.v. *Cape* is over 7,000 words long.

In many cases, users may not be interested in all available data, but only in parts of the entry. They should thus not be overloaded with information, but be able to access exactly the kind and amount of content they need for their purpose: they need selective access and selective data display (cf. Tarp 2012). The possibility of selectively querying the dictionary (along with optional access to desired components of full entries) seems to be a relevant means of ensuring customizable access to the dictionary. This will in turn be made possible by means of full markup of lexicographic data categories and appropriate user interface design.

Selective querying and data display is so far rarely offered in scholarly dictionaries: the *Trésor de la langue française informatisé* (*TLFi*) has the option to highlight lexicographic data categories in different colours; the South African *Woordeboek van die Afrikaanse Taal* (*WAT*), a multi-volume scholarly dictionary, only offers very limited possibilities, in its online version, for customizing entries (cf. Du Plessis 2014). The *OED* does offer category-based browsing and searching, as well as an internal vertical scrollbar which provides a useful signposting system ("how far down am I in this lengthy entry?") and an option to hide quotations. However, the entry microstructure, surrounding layout components (macrostructure) and navigation features rely on almost-exclusively textual presentations of a great deal of detail, resulting in a crowded interface which may deter some users and which is not optimised for mobile devices. In the online adaptation of the *DSAE* we aim to go beyond current strategies for selective querying and data display by meeting user demands for a simple user

interface which uses a layered approach to the presentation of detail and introduces interactive visual components to the historical dictionary model.

5 Initial Adaptation of Entry Display

At the time of the publication of the 2014 pilot edition, the structure of the XML dataset did not support flexible presentation of the entry layout. As such, the layout changes from print to the online edition were negligible, namely: the separation of the etymology (labelled ORIGIN) from preceding entry components, the positioning of definitions on a new line, and the hyperlinking of cross-references. The rest of the 2014 entry design closely resembles that of the print edition, as shown in Figure 2 below.

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**.

Origin: 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.*

Figure 2: Sample entry *aardvark* showing pre-adaptation layout of main entry components excluding quotations.

Alongside adaptations to the dataset to be published in 2016, an improved model for the entry layout and microstructure has been developed, as shown in Figure 3.

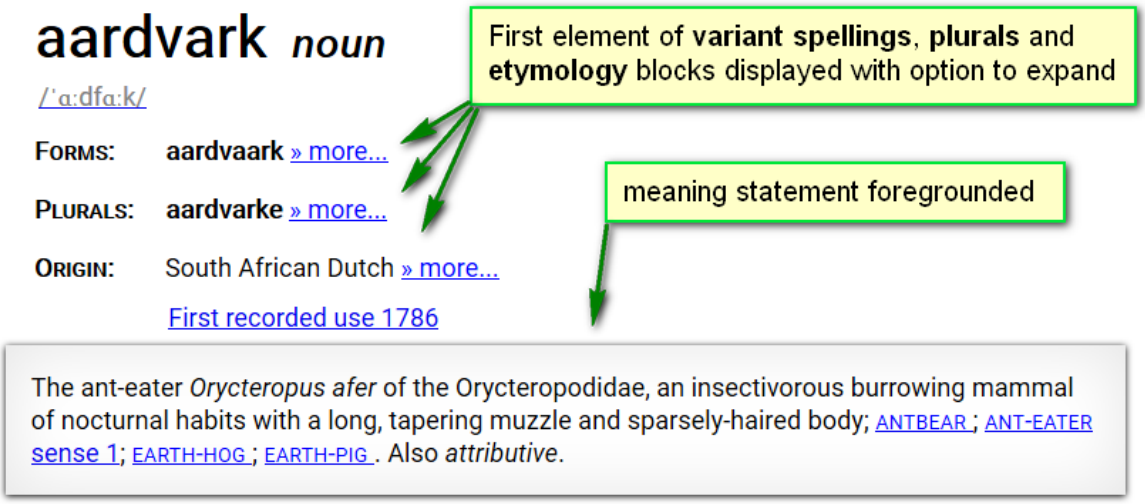


Figure 3: Wireframe model of adapted entry layout (excluding quotations).

The adapted layout displays the entry data in a way that is better suited to online presentation. Changes include: the separation and labelling of main entry components (forms, plurals and origin) into block-

level elements with optional expansion; the expansion of abbreviations and the inclusion of full forms for plurals, instead of *-s*, *-e* and *-en* (partly required to facilitate dictionary searches and search engine optimisation); and the foregrounding of data which are of immediate interest to most users, e.g. the meaning explanation.

The presentational adaptations illustrated in Figure 3 will be made possible through a subset of enhancements to the XML dataset which are nearing completion as of May 2016[2]. They are shown here as necessary background to the broader print-to-online adaptation project and specifically to the discussion of selective data display below.

6 Presenting Cognitively Relevant Data Online: the Case of Quotations

In the following, we give an example of an ongoing design study on the user-centred presentation of lexicographic data in the online *DSAE*, focusing on quotations and lemma dating.

6.1 Available Data

DSAE quotations all come with a publication year as well as a full bibliographic reference; original font styling and punctuation in quoted examples are reproduced, since they often reflect assimilation into SAE, showing the authors' attitude to the word in question (cf. examples (6) where italicisation signals self-conscious usage in a comparatively early context, and (7) where roman type suggests acceptance as a SAE term):

(6) *aard-wolf*: *ca.***1936** *S. & E. Afr. Yr Bk & Guide* 1072 The *Aard-wolf* (*proteles cristatus*), often called 'maanhaar jackal', stands about 18 inches in height and is of a brown colour with a few dark markings.
(7) *aardwolf*: **1982** *S. Afr. Panorama* Jan. 39 Also extremely rare is the aardwolf whose continued existence is dubious.

6.2 User Needs

The current version of the *DSAE* has quotation lists for each sense, sorted by publication date from the earliest to most recent attestation. These lists may be long. On average, each entry has about 10 quotations (with an average date span of 83 years). The quotations are integral to the dictionary design and serve several purposes. Their functions include showing example sentences, which sometimes elaborate on meaning; documenting first use and historical range per word sense; recording variant spelling patterns in their historical contexts; and, as indicated in 4.1 above, reflecting the degree of assimilation of loan words.

Given the historical emphasis of the dictionary and the fact that it documents only the South African variety of English, not general English, it is not offered as a general lookup tool to be consulted for

[2] Adaptations to the presentation of key components of the article layout and microstucture, and the dictionary macrostructure, are the subject of further collaboration with the University of Stellenbosch (see Du Plessis and Van Niekerk 2016).

communicative purposes. This market is already served by general English online dictionary products designed for that purpose (however SAE is necessarily given comparatively limited treatment in these reference works). In serving users wishing to consult an authoritative diachronic variety dictionary, on the other hand, the historical data need to be presented in more easily accessible ways. With respect to quotations and dating in the *DSAE*, the following user needs can be anticipated:

- to understand whether a given item is still current;
- look for the first attestation of an item;
- see attestations from a given period only;
- see the attestations presented in a different order (e.g. with chronology reversed);
- see the distribution of the *DSAE's* quotations over the timespan of South African English;
- not see any quotations at first, since the main interest is in definitions.

All the above are technically feasible, albeit with different kinds of impact on the dictionary design and requiring different sorts of automatic or manual changes to the XML dataset and online content display.

From March to April 2016 a user questionnaire was widely circulated requesting feedback on potential changes to the dictionary (see Van Niekerk *et al*. 2016). 70 respondents completed the survey. While this represents a fraction of the users who visit the dictionary website and more responses would have been desirable, the survey results give useful insights into trends among user preferences. On the topic of quotations, 5% indicated that they would like to see only the first recorded use, 3% that they would want to see only the first and last quotations, and 25% that they "are happy seeing the full list of quotations". The remaining 67% all preferred some form of optional display of citations, with 47% indicating that they would prefer the quotations to be hidden until a button or label was clicked to show them.

6.3 Optional Display of Quotations

Selective and optional display of quotations using simple, stable presentation devices is the most desirable solution to the challenge of meeting disparate user needs with respect to the amount of citation data being displayed. One possible approach to this problem could be to develop functionality which allows for specific user profiles offering different displays based on profile selection. While this would require rather complex changes to the interface and interaction mode of the *DSAE,* it is also not yet clear how many users would wish to engage with the time commitment of creating a user account and associated profile configuration. Our goal is therefore to meet most users' needs by providing the greatest possible ease of use and simplicity using the resources currently available.

25% of the questionnaire respondents indicated a preference for all quotations to be shown, while 67% preferred various forms of optional display of quotations. Displaying the full list of quotations for each entry as in the pilot edition would quickly lead to the user being overwhelmed with textual content, however, given the average count of 10 quotations per entry. This problem is of course amplified elsewhere in the dictionary: 5% of entries contain more than 25 quotations, and several high-frequency lemmas with nested compound forms contain over 100 quotations (e.g. *veld*, *boer*, *trek* and *Cape*, the latter containing 230).

As a simple means of satisfying most users' needs as indicated above, we decided on a content filtering

strategy which displays, by default, the first and last quotation per sense, with the option to click a button to expand to the full list. Figure 4 shows the resulting layout:

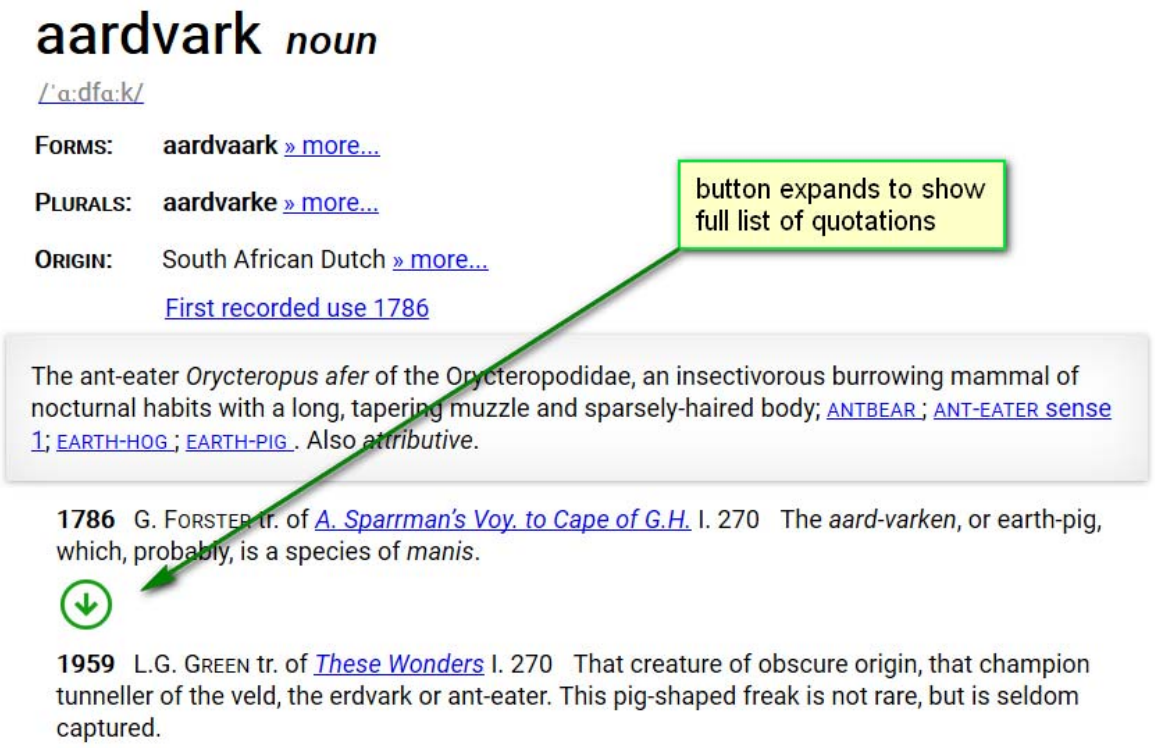


Figure 4: Wireframe model of entry *aardvark* displaying first and last quotation with the option to expand to the full list of 16 questions.

6.4 Visual Presentation Devices for Quotation Data

The selective and optional display of quotations discussed so far relates to the textual presentation of historical data. In order to mentally process the material presented, the user is still required to read through the excerpts and their condensed bibliographical headers in order to grasp important facts relating to historical context and distribution of documented usage across senses. Given that these data are encoded in machine-readable ways in the underlying XML markup for each entry, the relevant information can be presented more accessibly, namely through visual representations.

6.4.1 Time Bar Showing Historical Range

Although the *DSAE* provides citations of terms as far back as 1589 (e.g. *Cape*), these early examples often represent indirect evidence of usage and SAE terms only really begin to be frequently documented from the late 17th Century onwards. To give the user an instant indication of the historical span of each term, we have introduced a 'quotation bar' or 'time bar' component to the entry layout, as illustrated in Figure 5.

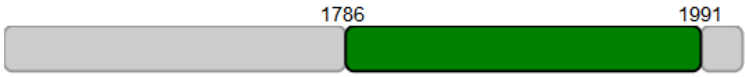


Figure 5: Time bar showing the range of quotations of *aardvark* (coloured green) in proportion to the timespan of South African English (grey).

The time bar is not a corpus frequency indicator. It is dynamically generated from the underlying quotation markup and is offered as a standard feature of each entry providing, at a glance, the historical span of quotations. In the case of *aardvark*, as illustrated, which contains only one sense, a single time bar suffices for the whole entry. Since quotation date spans vary in polysemous entries, however, a time bar is provided for each sense.

6.4.2 Histogram Showing Quotation Distribution

The time bar reduces the historical content of the entry model to a single axis, namely date range. A histogram offers a slightly more detailed, yet still simplified, visual representation of attestations by adding a second axis, namely the number of quotations in the article, per time interval. See Figure 6.

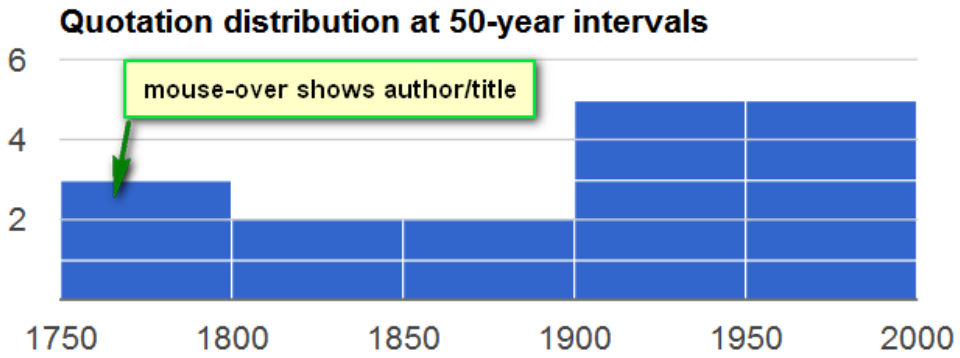


Figure 6: Histogram display of quotations for entry *aardvark*.

By introducing the histogram as a new visual component of the entry model we offer the user a quick overview of the distribution of citations, indicating documented usage at certain (here, 50-year) intervals. A single histogram is dynamically generated for each entry. Each coloured block represents a single quotation. Hovering the mouse pointer over a specific block causes the author or title to be displayed and the device is interactive insofar as each block is hyperlinked to the quotation it represents.

7 Explicating Data Categories in the *DSAE*

It is commonly accepted in e-lexicography that only a highly structured data collection offers lexicographers enough flexibility for customization of the data provided to users (cf. e.g. Bothma 2011). An implementation of a thoroughly user-centred approach to electronic dictionary-making thus requires a lexicographically structured data collection (called a database by some lexicographers) and a

set of filters used to determine the selection and ordering of data categories depending on user needs and users' pre-existing knowledge, as well as presentation filters to determine the wording, layout and display of lexicographic data (e.g. names of metalinguistic indicators, fonts, colours, etc. to be used).

7.1 Existing Data

As with many legacy dictionary datasets, not all lexicographic data categories were fully specified and marked up in the XML dataset, which is itself derived from an early SGML version of the dictionary created solely for print typesetting purposes. We have thus needed to further subdivide certain data categories, e.g. the element "label" <la>, which was used generically for several kinds of diasystematic marks. Furthermore, in some cases, sections of text with variable content had been grouped together instead of being subdivided into their different data categories: this was the case with meaning explanations, etymologies and cross-references to synonyms and related items. The practical result of this was that, in the early XML dataset, it was not possible to separate or filter adjacent entry components or to display them in different ways. Moreover, some entry components which could (and, in online displays, probably should) be presented as block-level components had been displayed as running text inseparable from neighbouring content: e.g. nested derivatives and compounds. Finally, not all items from specialised domains are annotated with an appropriate diasystemic label; we have addressed this issue with fauna and flora entries.

7.1.1 Procedures and Results

All our explication procedures are semi-automatic: pattern-based candidate identification and explication using global text processing techniques, followed by manual verification. We use indicators for data categories and/or for their start or end points in the entry text: in this way, overly generic markup or clustered data types were distinguished and separated into more usefully semantically-tagged values. An example case of label markup is given below.

7.1.1.1 Explicating Lexicographic Data Categories: the case of <la> ("label")

The dictionary had a range of labels, in three general categories:

- **Register labels**
  - e.g. <la>jocular</la>, <la>Army slang</la>
- **Provenance labels**
  - e.g. "Used also in <la>British English</la>."
- **Subject labels**
  - e.g. <la>Law</la>, <la>Music</la>, <la>Farming</la>

These categories were not distinguished in the original markup; all simply used the same generic <la/> element. Label categories have now been analysed and their markup enhanced to show their categories, namely:

- **Register**: <la type="**register**">jocular</la>
- **Provenance**: <la type="**provenance**">British English</la>

- **Subject**: <la type="**subject**">Architecture</la>

Using the methods described above, over 3,500 instances of the element <la> were automatically split into grammatical, style/register and domain labels after manual classification of the existing values. Similarly, variant spellings, plural forms, nested compounds and derivatives – all formerly presented as flowing text in unwieldy line-wrapped paragraphs – have been resolved into a single block-level element for each component. This groundwork opened the way to selective and optional presentation of entry content, which will also facilitate further adaptations in future.

The abovementioned subcategorisation of labels serves another purpose in that it makes label values machine-readable and therefore available for selective querying. For example, Architecture, Law or Music labels can act as subject category indicators, allowing the relevant lemmas to be filtered in combination with other values (e.g. to show "Music terms derived from Malay") or else linked to one another (e.g. "Law terms derived from Afrikaans"). Subject labels were however not originally intended to perform the role of subject category metadata and, while they aid lemma categorisation, the use of subject labels is limited in proportion to the whole dictionary and they therefore cannot be used on their own. Subject classification requires a global strategy to cover all senses, and this topic is discussed in the next section.

7.1.1.2 Subject Categorisation of Senses: the example of Fauna and Flora

Due to its wide historical range and its emphasis on comprehensive coverage of a language variety, the *DSAE* has an inclusive policy with respect to names of local plants and animals. This subset of lemmas was semi-automatically classified in an initial phase of sense categorisation. Since their definitions specified scientific names, it was possible to detect flora and fauna items based on morphological characteristics of botanical and zoological nomenclature and by matching against scientific spellchecker dictionaries made available by the South African National Biodiversity Institute (see Powrie 2013). The procedure was as follows, using custom software developed for the purpose:

- using pattern matching, extract words from definitions matching the morphology of taxonomic names (order, family, genus, or Latin binomial or trinomial species or subspecies names);
- compare these values with the scientific names listed in the botanical and zoological spellchecker dictionaries;
- when matches are found, automatically apply appropriate subject category markup to the *DSAE* values.

The scientific spellchecker dictionaries were found to be comprehensive and reliable and only a small number of failed matches (due to misspellings, upper-/lowercase variations or outdated scientific names in the *DSAE*) required manual correction and categorisation. This exercise resulted in the classification of over 2000 nested species descriptors within 794 *DSAE* entries, amounting to 17% of the dictionary. Given that languages of origin are already encoded in the etymology components of all entries, this facilitates selective querying for the kind of use case posed in section 2.1 above, namely to answer the question "Which plant names in South African English derive from Khoikhoi?"

Subject categorisation of the remaining senses is, however, a largely manual task and is a focus area of current work.

7.2 Status as of May 2016

The markup enhancement is ongoing; while the above-mentioned data categories have already been handled as described, more work is being undertaken to add subject classification to all senses, isolate meaning explanations (which follow several different models) and to add specificity to cross-reference hyperlinks or remove them where they can be replaced with their target content.

8 Conclusion and Future Work

This article describes work undertaken on the further adaptation of the 2014 pilot edition of the online *DSAE*, along two main development paths: (1) the enhancement of the dictionary dataset to support redesigned entry presentation for electronic platforms, introducing optional and selective display strategies; and (2) the introduction of new visual components to a previously exclusively textual entry layout, to make it easier for users to interpret information presented in condensed form in historical dictionary entries.

To achieve this, a number of lexicographic data categories have been explicated and marked up or still need to be in future; this work is ongoing. While we have started to design new ways of presenting cognitively relevant data with respect to quotations and lemma dating, we will develop more options for both text reception use as well as for different types of cognitive uses. In particular, we aim to make selective cross-dictionary queries possible (cf. examples 4 and 5 in section 2.1). Current work also lays the groundwork for a future, mobile-optimised interface for the *DSAE*.

9 References

Bothma, T.J.D. (2011). Filtering and adapting data and information in an online environment in response to user needs. In Fuertes-Olivera, P.A. & Bergenholtz, H. (Eds.), *e-Lexicography. The internet, digital initiatives and lexicography*. London: Continuum, pp 71–102.

Du Plessis, A. (2014.) A functional analysis of the e-WAT with specific focus on the mobile version: towards a model for improvement. In *Lexikos* (24), pp 75–93.

Du Plessis, A. & Van Niekerk, T. (2016). Redesigning the DSAE's presentation and navigation features: adapting a historical dictionary for the modern online user. 21st Afrilex Conference, 4-7 July 2016 (forthcoming conference presentation). Tzaneen, South Africa.

Powrie, L.W. (2013). A list of South African botanical names for spell checking. South African National Biodiversity Institute, Cape Town. Accessed at: http://www.sanbi.org/information / infobases/biodiversity-research-tools [01/11/2015].

Powrie, L.W. 2013. A list of South African zoological and other (including fungi and lichen) names for spell checking. Souths African National Biodiversity Institute, Cape Town. Accessed at: http://www.sanbi.org/information/infobases/biodiversity-research-tools [01/11/2015].

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

Dictionary of South African English Online (pilot edition, 2014; current edition, 2016). Accessed at: http://dsae.co.za [30/04/2016].

Tarp, S. (2008). Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography. In *Lexicographica*. Series Maior 134. Tübingen: Max Niemeyer.

Tarp, S. (2012). Online dictionaries: today and tomorrow. In *Lexicographica*, (28), pp 253–267.
Van Niekerk, T., Morris, L., Wolvaardt, J., & Le Du, B. (2016). User survey: Dictionary of South African English Online. Dictionary Unit for South African English.