Semi-automating the Reading Programme for a Historical Dictionary Project

Tim van Niekerk,
Lexicography Unit for South African English
at Rhodes University, Grahamstown EC, South Africa
and
Johannes Schäfer, Heike Stadler and Ulrich Heid,
Universität Hildesheim, IwiSt-STCL, Germany

Euralex-2018, Ljubljana, July 2018

Overview

- DSAE: characterization and history,
  macrostructure and microstructure
- Objectives of the project
- Data sources: creation of a corpus of SAE
- Tools for dictionary updating: Scenario and results
  - Search for quotations
  - Search for spelling variants
  - Search for inclusion candidates
- Conclusion and future work

Acknowledgements....

- ...to the DSAE Lexicography Unit
  - for sharing DSAE dictionary data with University of Hildesheim
  - for allowing Tim van Niekerk a study leave to Hildesheim
- ...to the Niedersächsische Staatskanzlei
  and the Ministry of Science and the Arts Niedersachsen (MWK)
  for a second round of financial support, in 2017
- ...to the *Wortschatz* project group at University of Leipzig
  - for making their corpus data available
  - for allowing us to process these data

DSAE

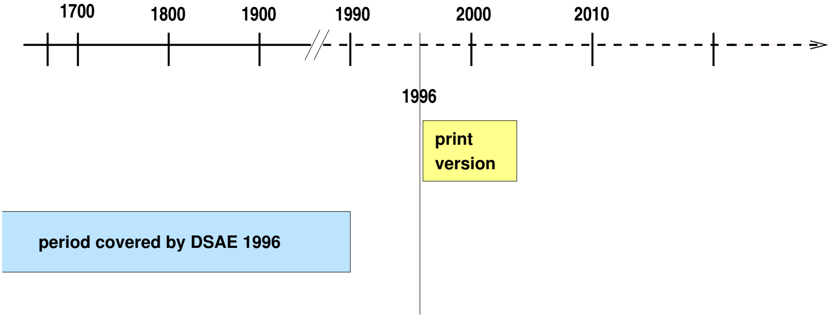
A short characterization

- DSAE:
  Dictionary of South African English on Historical Principles, 1996
  - OED-style diachronic variety dictionary
  - Printed version:
    - 850 pages, monovolume scholarly dictionary
    - Variety dictionary with definitions and quotations
  - Major information source on South African English (=SAE)

DSAE's history (1/3)

Publication dates and lexical coverage

- Printed dictionary published in 1996 —
  covering data from early times of SAE to 1990

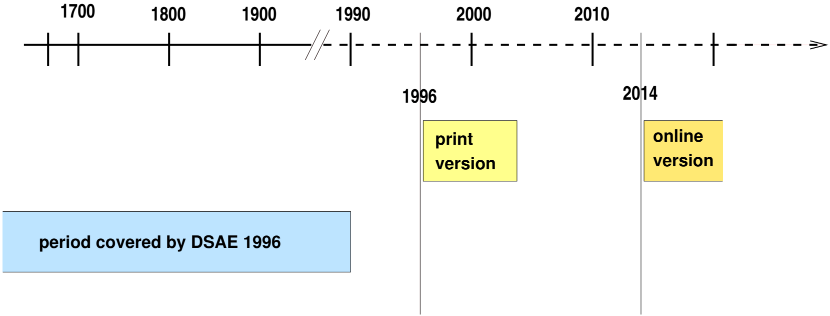


DSAE's history (2/3)

Publication dates and lexical coverage

- Printed dictionary published in 1996 —
  covering data from early times of SAE to 1990
- Out of stock from 2004 onwards,
  online version since 2014

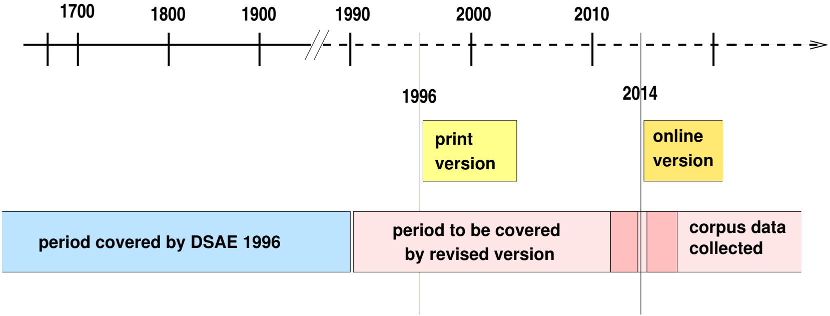
http://dsae.co.za



DSAE's history (3/3)

Publication dates and lexical coverage

- Printed dictionary published in 1996 —
  covering data from early times of SAE to 1990
- Out of stock from 2004 onwards,
  online version since 2014 http://dsae.co.za
- Revision planned and partly under way —
  Experimental corpus creation for 2011-2017



Principles underlying the DSAE

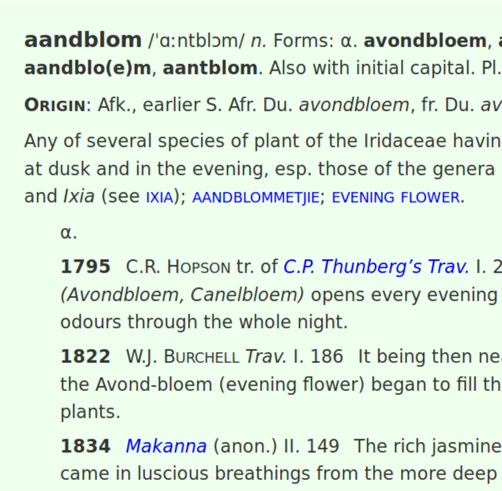
Macrostructure

- 4,600 lemmas
  plus ca. 6,000 orthographic variants of these lemmas
- Plus ca. 4,000 derivatives and compounds

Principles underlying the DSAE

Microstructure

- Data categories (1/2)
  - Lemma sign
  - Orthographic variants
  - Inflection forms
  - Grammar + pronunciation
  - Etymology,
    Word history
  - Diasystematic marks:
    domain, usage, ...
  - Meaning paraphrases,
    synonyms



aandblom /ˈɑːntblɔm/ *n.* Forms: α. **avondbloem**, **a**
aandblo(e)m, **aantblom**. Also with initial capital. Pl.

Origin: Afk., earlier S. Afr. Du. *avondbloem*, fr. Du. *av*

Any of several species of plant of the Iridaceae havin
at dusk and in the evening, esp. those of the genera
and *Ixia* (see IXIA); AANDBLOMMETJIE; EVENING FLOWER.

α.

1795 C.R. Hopson tr. of *C.P. Thunberg's Trav.* I. 2
(Avondbloem, Canelbloem) opens every evening
odours through the whole night.

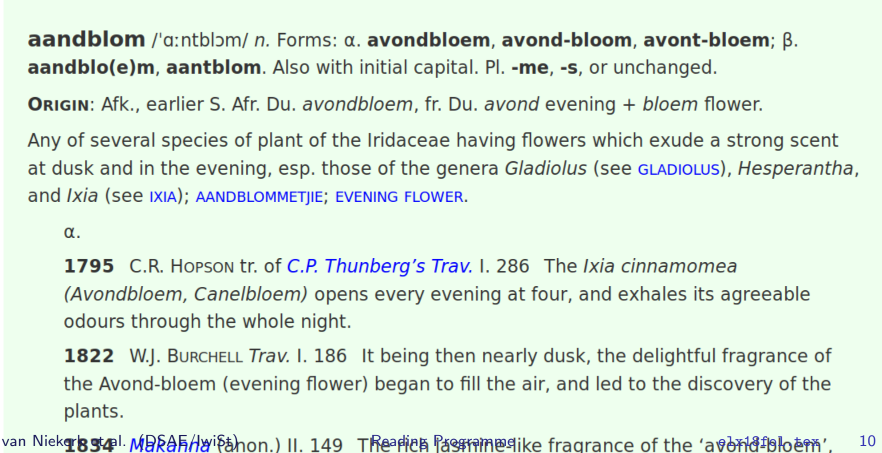
1822 W.J. Burchell *Trav.* I. 186 It being then ne
the Avond-bloem (evening flower) began to fill th
plants.

1834 *Makanna* (anon.) II. 149 The rich jasmine
came in luscious breathings from the more deep

Principles underlying the DSAE

Microstructure

- Data categories (2/2)
  - 45,00 quotations with full bibliography:
    author, date, references of publication cited
- Quotations derived from ca. 300,000 index card citations



aandblom /ˈɑːntblɔm/ *n.* Forms: α. **avondbloem**, **avond-bloom**, **avont-bloem**; β. **aandblo(e)m**, **aantblom**. Also with initial capital. Pl. **-me**, **-s**, or unchanged.

ORIGIN: Afk., earlier S. Afr. Du. *avondbloem*, fr. Du. *avond* evening + *bloem* flower.

Any of several species of plant of the Iridaceae having flowers which exude a strong scent at dusk and in the evening, esp. those of the genera *Gladiolus* (see GLADIOLUS), *Hesperantha*, and *Ixia* (see IXIA); AANDBLOMMETJIE; EVENING FLOWER.

α.

1795 C.R. HOPSON tr. of *C.P. Thunberg's Trav.* I. 286 The *Ixia cinnamomea (Avondbloem, Canelbloem)* opens every evening at four, and exhales its agreeable odours through the whole night.

1822 W.J. BURCHELL *Trav.* I. 186 It being then nearly dusk, the delightful fragrance of the Avond-bloem (evening flower) began to fill the air, and led to the discovery of the plants.

1834 *Makanna* (anon.) II. 149 The rich jasmine-like fragrance of the 'avond-bloem',

van Niekerk et al. (DSAE/IwiSt) Reading Programme elx18fel.tex 10 / 29

Objectives of the project

Experiments towards a semi-automatic reading programme

- Corpus creation for 2011-2017,
  as a source for extracting raw material for DSAE entries
- Creation and application of tools
  - Finding new quotations for lemmas from DSAE,
    covering all known orthographic variants
  - Finding new orthographic variants of lemmas from DSAE
  - Finding new lemma candidates
    and quotations to document them

Data sources

Corpora containing South African English (1/3)

- Source 1: News corpus 2015-2017 work by H. Stadler
  Collection of online newspapers

BusinessLive	Economy, Politics, Industry
Sowetan	National general newspaper
TimesLive	National general newspaper
Dispatch	Regional newspaper, Eastern Cape
Citizen	Tabloid newspaper, regional: Johannesburg
Daily Maverick	Opinion-oriented, partly satirical

  - Collected daily, documented with metadata
  - Ca. 100 million words

Data sources

Corpora containing South African English (2/3)

- Source 2: Generic web corpus created by the Leipzig CURL project: *Crawling under-resourced languages* Quasthoff et al. 2015
  - Domain .za
  - Sentence-wise, sentences scrambled
  - Limited bibliographical or source metadata
  - Ca. **3 billion words**

Data sources

Corpora: Metadata available (3/3)

- News corpus (100 M)
  - Name of website
  - Date of publication, date of crawling
  - Newspaper section
  - Name of author (if present)
- Web corpus (3 B)
  - Name of website
  - Date of crawling

Data sources

Computational linguistic processing of the corpora

- Annotation

POS tagging	Treetagger Multext Tagset	Schmid 1994
Lemmatization	TreeTagger +Lexicon	

- Metadata annotation (as far as available)
- Preparation for query

Encoding	CWB: Open Corpus Workbench	Evert/Hardie 2011

Data sources

Annotated corpus data

Sample sentence for the item *aandblom*

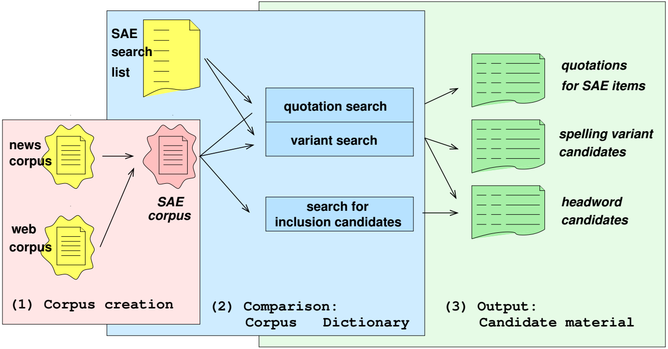
```
The/DT/the
delicate/JJ/delicate
Hesperantha/NP/Hesperantha
cucullata/NNS/cucullata

(/(/(
aandblom/NN/aandblom
)/)/)

is/VBZ/be
a/DT/a
geophyte/NN/geophyte
with/IN/with
sword-shaped/JJ/sword-shaped
leaves/NNS/leaf
,/,/,

growing/VBG/grow
15/CD/@card@
-/:/-
30/CD/@card@
cm/NN/cm
tall/JJ/tall
./SENT/.
```

Tool scenario

Major steps

(1) Corpus creation: news $\oplus$ web corpus → SAE corpus

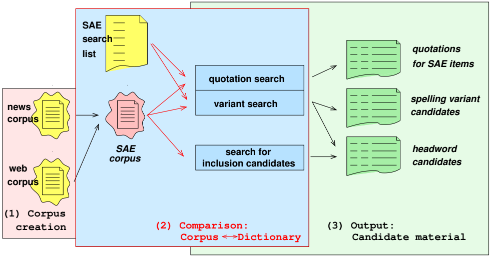
(2) Comparison between DSAE dictionary and SAE corpus
Preliminary step: annotation of names (lexicon-based)



Tool scenario

Comparison steps (1/2)

- Input:
  DSAE search list = lemmas $\oplus$ variants $\oplus$ inflected forms
- Quotation search:
  Sentences from the SAE corpus where search list items appear
- Variant search:
  Forms unknown to DSAE and TreeTagger, sorted by edit distance

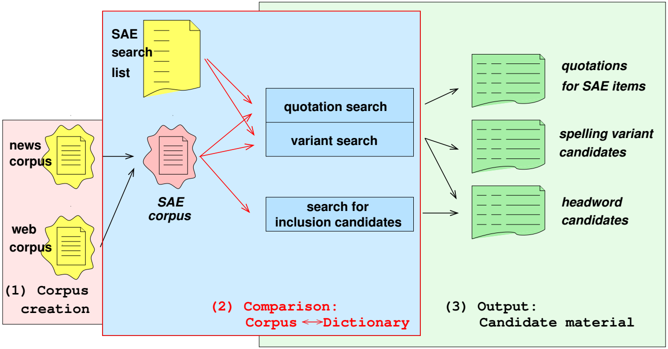


Tool scenario

Comparison steps (2/2)

- Search for inclusion candidates: Comparison of SAE corpus items with BNC: listing items that are neither in DSAE nor in BNC, and which are not covered by variant search

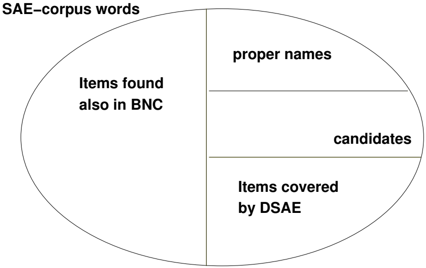
Ahmad et al. 1992



Using the SAE corpus

Comparison steps and SAE corpus subsets

- Subsets of the lexical inventory of the SAE corpus
  - (a) Items contained in BNC
  - (b) Proper names
  - (c) Items covered by the DSAE dictionary
  - (d) Potential candidates



- Comparison steps
  - Quotation search: Example sentences from (c)
  - Variants: (c) ↔ (d)
  - New items: (d) vs. BNC

- All steps require manual inspection!

Results: Quotation search

Output features – an example

- Quoted text is annotated with provenience metadata
- Quotation is linked to headword from DSAE
- Example: *aandblom*

```
<dictionary date="21-12-2017 22:37">
<dictionary_id id="e00005">
<dictionary_entry frequency="5"
                  string="aandblom"
                  type="[headword]">
<example_sentence cwb_corpus_name="SAE2"
                  doc_date="2014-09-21"
                  doc_id="1415967"
                  document="http://www.venturesintoafrica.co.za/"
                  p_id="26226938"
                  subcorpus="eng-za_web_2014">

How else will you know that a Babiana (Bobbejaantjie) smells
like baby talcum and the Hesperantha (aandblom) smells heavenly ?
                                                 </example_sentence>
```

Results: Quotation search

Quantitative aspects

- DSAE headwords and derivatives/compounds covered:
  7,025 items with a total of 21,768 variants
- Quotations available for ca. 85% of the headwords
- For around 50% of the items,
  the SAE corpus contains ca. 100 candidate sentences

Results: Variant search

Preparation of dictionary data

- DSAE contains headwords and variant types
- From these, hypothetical variants are generated, e.g. by inserting or leaving out hyphens or blanks
- Example:

```
e00005      [dictionary ID]
aandblom    [headword]
avondbloem  [variant spelling]
avond-bloom [variant spelling]
avondbloom  [variant spelling] [variant spelling generated]
avond bloom [variant spelling] [variant spelling generated]
avont-bloem [variant spelling]
avontbloem  [variant spelling] [variant spelling generated]
avont bloem [variant spelling] [variant spelling generated]
aandbloem   [variant spelling]
aandblom    [variant spelling]
aantblom    [variant spelling]
aandblomme  [plural of headword]
aandbloms   [plural of headword]
```

Results: variant search

Overview – example

- Items from the corpus are checked in terms of their edit distance Levenshtein distance
- Candidates are shown with absolute frequency and distance measure
- Example: *imphepho* (a medicinal plant)

```
imphepho      [catchword]

[iphepho, 3, 1] [mphepho, 4, 1] [imphephu, 3, 1] [impepho, 77, 1] [imphepo, 13, 1]
[iphepha, 15, 2] [mpepho, 16, 2] [mphephu, 5, 2] [iphupho, 5, 2] [kuphephe, 2, 3]
[umkhapho, 4, 3] [umthetho, 20, 3] [imperio, 5, 3] [iPhepha, 14, 3] [amaphepha, 15,
```

Results: variant search

Heuristics and methods of result analysis

- Shorter search terms: distance 2-3,
  longer search terms: distance 4-5
- Example: *karretjie people*
  (a nomadic people who travel in animal-drawn carts)

```
Search term: karretjie people

Variants:    karretjiepeople       | without space
             karretjie-people      | with hyphen

             karretjiemense        | Afrikaans compound
             karretjie-mense       | dito, hyphenated (by EN-speaking author)
             karretjiesmense       | with Afrikaans-style plural marker

[karretjiemense, 39,5] [karretjie-mense, 4,5] [karretjiesmense, 4,5]
```

→ all forms in *-mense* seem to be in use

Results: search for inclusion candidates

As a side-effect of variant search

- Items thrown up by Levenshtein comparison
  with high frequency,
  which are not related to the search terms
- Example:
  - Start with *bogadi* (traditional African wedding gift)
  - find:
    moladi
    (system for rapid and inexpensive wall building using precast moulds)

```
bogadi    [headword]    [bosadi, 8, 1] [baladi, 10, 2] [bogi, 4, 2]
[boga, 28, 2] [Dogade, 2, 2] [boondi, 3, 2] [nogado, 2, 2] [bhadi, 3, 2]
[bovada, 3, 2] [Jogami, 18, 2] [bogart, 5, 2]
[moladi, 120, 2]
```

Results: search for inclusion candidates

Use of traditional term extraction techniques

- Comparison of relative frequency: Ahmad et al. 1992
  SAE corpus ↔ BNC, for each item
- Keep those items that are much more prominent in SAE
- Examples found:

-*braairoom*	entertainment room used for indoor barbecues
-*mokoro*	type of canoe used in Botswana
-*miombo*	a Southern African vegetation type
miombo woodlands	

Conclusion

We have shown:

- how a corpus of news and web data was compiled
  that can be used for documenting entries of DSAE
- how spelling variants of SAE items can be found
- how raw material is found
  from where inclusion candidates for DSAE can be gathered

- how the above can be achieved
  with standard computational linguistic corpus technology
- that all results need to be further checked by experts

Future work

Planned steps

- Further exploration of the data accumulated: qualitative (and to some extent quantitative) evaluation
- More devices to reduce the amount of data presented to the lexicographer
  - e.g. filters by frequency
  - e.g. “goodex”-like tools to select quotations
- Making the workflow directly usable for lexicographers, e.g. on new text data