Henning Bergenholtz[1, 2 & 3]

Theo JD Bothma[2]

Rufus Gouws[3]

[1]Centre for Lexicography, University of Aarhus
[2] Department of Information Science, University of Pretoria
[3] Department of Afrikaans and Dutch, Stellenbosch University

A model for integrated dictionaries of fixed expressions

eLEX2011, Bled, Slovenia
10 – 12 November 2011

AARHUS UNIVERSITY

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Overview

- Contemplative vs transformative
- Danish dictionaries of idioms / fixed expressions
- Afrikaans dictionaries of idioms / fixed expressions
- A database for Afrikaans fixed expressions
- One database, six dictionaries
  - A database is not a dictionary
  - A dictionary is not a database
- Forthcoming attractions
- Conclusion

Contemplative vs transformative

- Contemplative
  - Analysis of existing dictionaries
- Transformative
  - Theoretical analyses
    - Potential user situations
    - Respective user conditions
    - User needs
  - Used to develop new concepts for compiling new dictionaries
    - Typically monofunctional dictionaries
  - On the basis of theoretical analyses the lexicographer decides what the characteristics are of the mono-functional dictionaries that will satisfy specific user needs

Danish dictionaries of idioms

- Analyses of
  - User feedback through e-mail
  - Log files
- Number of searches in *Den danske Idiomordbog*

With result	70.4%
Without result	29.6%

- Users were looking for idioms that were in fact not idioms
- Don’t compile a dictionary of idioms
- Widen scope to include all types of fixed expressions
- Rationale for Afrikaans database

Afrikaans dictionaries of fixed expressions

- Existing Afrikaans dictionaries
  - Five restricted dictionaries from 1924 to 2009
  - Various general monolingual and bilingual dictionaries offer an extensive presentation of fixed expressions
- Shared feature
  - All in printed format
- Some general monolingual and bilingual dictionaries are available on CD ROM and online
- No bad e-dictionaries of fixed expressions
- Transformative approach to planning and compilation

Database for Afrikaans fixed expressions

- Based on the Danish experience of their database of fixed expressions
  - Danish database 14 fields
  - Afrikaans database 36 fields
- Database developed in MySQL
- Open source software technologies
  - HTML, XML, XMLT, Perl, CGI and related technologies

Database for Afrikaans fixed expressions (2)

- Database management system
  - Comprehensive administrative back-end
    - Manages access, data security and integrity, version control and back-up
  - Two further interfaces
    - For researchers
    - For end-users

Database fields

1. Core field
2. Meaning in Afrikaans
3. Internet link to meaning
4. Further meaning item in Afrikaans
5. Meaning in English
6. Grammar
7. Comment on grammar
8. Internet link to grammar
9. Background remark(s)
10. Comment on background remark(s)
11. Internet link to background remark(s)
12. Fixed expression(s) in Afrikaans
13. Remarks on the fixed expression(s)
14. References to fixed expression(s)
15. Internet link to variants, e.g. statistical
16. Fixed expression(s) in English translated from Afrikaans
17. Style
18. Comment on style
19. Internet link to style
20. Classification of the fixed expression
21. Comment on classification
22. Collocation(s)
23. Comment on collocation(s)
24. Internet link to collocation(s)
25. Example(s)
26. Comment on example(s)
27. Internet link to example(s)
28. Synonym(s)
29. Comment on synonym(s)
30. Internet link to synonym(s)
31. Antonym(s)
32. Comment on antonym(s)
33. Internet link to antonym(s)
34. Associated concept(s)
35. Key word(s)
36. Memo field

Comments on database fields

- Field 1
  - Core or lemma field
  - Used for
    - automatic searches
    - items the user can use as links if search results are displayed as a list, or if synonyms or antonyms are provided
  - The field contains key words with all the lexical words, including irregularly conjugated forms which occur in the fixed expression(s)

Comments on database fields (2)

- Field 9
  - A brief history behind the full expression
  - If there are two different histories and it is not clear which one is correct, both are given
  - Reference to background histories in various textbooks and dictionaries
- Field 12
  - Only one expression, or plus variants if they exist

Comments on database fields (3)

- Field 22
  - ‘Collocations’ in the sense of combinations of words in which the fixed expression occurs
  - A collocation is never a complete sentence
- Field 25
  - ‘Example’ refers to a full sentence
- Field 34
  - Associated concept(s)

One database, six dictionaries

- Meaning of Fixed Expressions
- Use of Fixed Expressions
- Fixed Expressions with a Specific Meaning
- Knowledge about Fixed Expressions
- Afrikaans-English Dictionary of fixed expressions
- Comprehensive Knowledge about Fixed Expressions

One database, six dictionaries (2)

- For each dictionary the search occurs in different fields
- Different search orders in fields
- Presentation of data from different fields for each dictionary
  - Each dictionary is unique in terms of the fields that are displayed to represent an article
- Only the dictionary “Comprehensive Knowledge about Fixed Expressions” displays all data
  - With exception of memo field
- See Proceedings for details of each dictionary

Forthcoming attractions

- A database of 10,000 to 15,000 cards
- Make available when only 1,000 cards are ready
- Amend or add specific / additional data on the basis of
  - log file analyses and user feedback
  - further research on and experimentation with concepts and tools for manipulating data in the e-environment
    - User profiling
    - Personalized search and display options
    - Additional fields for more detailed information
    - Multi-language databases

User profiling

- Users will be able to define a user profile at the beginning of a consultation session
- Set up a persistent profile
  - Remain active across multiple user sessions
  - Ability to reset or change this profile at any stage
- Profiles will enable users to
  - define the specific dictionary they intend consulting
  - set personalised search and display options

Personalized search and display options

- Six dictionaries are six different customised views on the database
  - Each is defined in terms of a specific type of user need defined by the lexicographer
- Provide the user with the option to define his/her own search
  - Define his/her own personalised / customised dictionary
- Customised advanced search and display facilities
  - The user can define exactly which data are to be displayed
  - Data of only a single field or any combination of fields

Additional fields for more detailed information

- Currently we assume that all users require the same amount of detail when accessing a dictionary article
- This is not necessarily the case
  - Some users may require only a brief description
  - Others may require a detailed exposition
- Data will be made accessible on demand
  - By means of a “Read more” button
  - By adapting the user profile at the start of the consultation session

Additional fields for more detailed information (2)

- Database structure makes provision for examples
  - Comments about and links to the original contexts
- Highly selective list of examples to illustrate meaning and use of a specific fixed expression
- In individual cases users may require either more examples or additional detail
  - Actual examples in context
    - Concordance of examples in a keyword in context (KWIC) format
  - A table that provides only a statistical analysis of the occurrence of variants at a specific time

Additional fields for more detailed information (3)

- Two different types of tool to
  - present “raw” corpus data in a KWIC format
  - do statistical analysis of the “raw” corpus data and present the results in statistical tables
- Research required
  - How should the data in the external database(s) be marked up to enable access to specific data at a fine level of granularity?
  - How are word form variants to be handled?
    - Detailed tagging of morphological forms beforehand
    - Link to the “raw” text of the corpora on the fly without prior tagging
  - What type of tools will be required to make this type of searching/linking possible

Multi-language databases

- Currently Afrikaans and only a single field for English
- Concepts and database structures can be used for other languages
- Feasible to create multiple interlinked databases for fixed expressions in multiple languages
  - Interlinked via a pivot language, for example English
- Existing databases could also be linked
  - Minimum requirement
    - There should at least be a minimum set of corresponding fields, or
    - Translation tables between different fields should be created

Conclusion

- Some of the envisaged expansions may not necessarily currently be commercially feasible
  - Time / cost required to do the programming or to write / collate / select the data may simply be too much
  - Some may not be what users may require
- If researchers do not experiment with concepts and technologies that currently do not seem commercially feasible, innovation will be stifled
- Such “blue sky” research could lead to e-information tools that are not only incrementally better, but provide different tools through disruptive innovation

Conclusion (2)

- Two aims of current project:
  - To create a database of fixed expressions, as well as to develop the necessary database tools, administrative backend, user interfaces and search functions
    - that enable users to have access to a number of dictionaries
    - to result in a useful product this has to be completed in a limited timeframe
  - To provide a platform to
    - experiment with disruptive technologies
    - see to what extent any of these technologies can
      - add value for the user in providing access to information
      - in terms of the user's specific information need in a given user situation

Conclusion (3)

- Users may help to improve e-dictionaries incrementally
- Only fundamental research in
  - meta-lexicography
  - user needs
  - database technologies
  - principles of information organisation, access and retrieval

  will result in different types of e-tools

Thank you!

Questions / comments?

hb@asb.dk

theo.bothma@up.ac.za

rhg@sun.ac.za