EURAC research

INSTITUT FÜR DEUTSCHE SPRACHE

The lexicographic working environment in theory and practice

Andrea Abel, EURAC - Bozen/Bolzano (Italy), andrea.abel@eurac.edu

Annette Klosa, IDS - Mannheim (Germany), klosa@ids-mannheim.de

EURALEX 2012
Oslo, 6-11 August 2012

Outline

1. Introduction
2. The modern lexicographic working environment
   - Main characteristics of a dictionary writing system
   - Off-the-shelf versus in-house software
3. An example: *elexiko*
   - The *elexiko* project
   - A critical reflection on the *elexiko* working environment
4. Conclusions and outlook

1. Introduction

- Growing automation & ensuing demands for targeted software tools → gap in meta-lexicographic research
- Aims of the talk:
  - to contribute to fill this gap (focus on academic non-commercial lexicography, with the help of a real-life example)
  - to reflect on selection criteria for lexicographic tools and to propose some recommendations for how to structure the lexicographic working environment

2. The modern lexicographic working environment

- Dictionary writing system (DWS):

  "[...] a piece of software for writing and producing a dictionary. It might include an editor, a database, a Web interface and various management tools [...]" (Kilgarriff 2006: 7)

- Examples:
  - commercial products (e.g. *ABBY Lingvo Content*, *IDM DPS*, *iLEX*, *TLex*)
  - in-house tools (e.g. *ANW Article Editor*, *DEB*, *EELex*, *Onoma*)

(cf. Abel in print, de Schryver 2011)

2.1 Main characteristics of a DWS

- Three main components:
  a) an editing tool
  b) a database
  c) administrative tools

  (cf. Atkins & Rundell 2008, Svensén 2009, de Schryver 2011, Abel in print)

2.1 Main characteristics of a DWS

a) The editing tool, e.g.:

- different visualization options, usually a ‘what-you-see-is-what-you-get’ (WYSIWYG) and a ‘tree-diagram’ view
- consistency of entries, e.g. limited number of values (e.g. grammar codes) for certain fields or use of drop-down lists
- automatic generation of some non-typographical structure indicators (e.g. commas, brackets)
- integrated style guide
- copy and paste functions

2.1 Main characteristics of a DWS

- use of ‘templates’ for typical entry-types or recurring parts of entries
- validation of the entry structure, i.e. check that the syntax corresponds to the dictionary’s DTD
- support with error-prone procedures, e.g. automatic re-numbering of senses, update of cross-references
- real-time spellchecker
- ...

2.1 Main characteristics of a DWS

b) The database, e.g.:

- complex searches, e.g. for entries written or modified by an editor
- use of Unicode that recognizes all characters
- a server-client architecture that easily allows distributed work via Internet from any location
- usually use of XML/DTD, XML Schema or own formats based on XML
- data import and export from and into different formats, e.g. XML, RTF, PDF, HTML
- ...

2.1 Main characteristics of a DWS

c) Administrative tools, e.g.:

- a ‘workflow manager’, e.g. to allocate batches of entries to be compiled to single lexicographers
- search options, e.g. to keep track of progress against the working schedule
- feedback possibilities to lexicographers
- assignement of read and write permissions
- lock of individual fields, e.g. that are to be edited by specialists

2.1 Main characteristics of a DWS

- version control
- automation of processes through scripts, e.g. mass data update
- batch merges
- ...

2.2 Off-the-shelf versus in-house software

- tendency:
  - commercial publishers (e.g. all the main English language publishers) → off-the-shelf packages
  - acadamic institutions (e.g. the *DWDS* or *elexiko* initiatives), but also some publishers (e.g. *DUDEN*) → in-house software
- existence of excellent, highly sophisticated off-the-shelf DWSs →
  development of an in-house tool = reinventing the wheel?
  (cf. de Schryver 2011)

2.2 Off-the-shelf versus in-house software

- Possible reasons for the development of own tools:
  - the development of own tools as a part of the reasearch activities of an institution (e.g. *Jibiki*, *elexiko*)
  - further development of home-grown systems, often in long-term projects, usually started before commercial packages were available
  - off-the-shelf tools do not fulfill the requirements of a specific project (e.g. *DUDEN-Wissensnetz*)
  - lack of knowledge about commercial products and their features → no informed decision

3.1 An example: *elexiko*

- Corpus-based, monolingual German online dictionary:
  - www.elexiko.de
  - 300 000 entries, dictionary under construction
  - published as one of the reference works in the dictionary portal *OWID* (www.owid.de)
  - work in progress since ca. 10 years

3.1 An example: *elexiko*

- Working environment: several components
  - *COSMAS II* as the corpus tool to process data from the *elexiko-corpus*
  - XML-editor: at first *XMetaL*, now *Oxygen*
  - cross-reference-management tool (*Vernetziko*), supporting the creation and maintenance of links between entries and other lexicographical processes

3.1 An example: *elexiko*

- Working environment: several components
  - *EDAS*-interface to the *ORACLE* database storing the data; offering search options; also used as a workflow manager
  - in-house presentation similar to web presentation
  - electronic dictionary manual and documentation of DTD
  - electronic secondary sources (other dictionaries, grammars)

3.1 An example: *elexiko*

- Workflow:
  - checking out an entry to be worked on through *EDAS*, locking it for other editors
  - searching the corpus, analyzing the corpus data with *COSMAS II*
  - editing the entry in *Oxygen*, including citations from the corpus via copy & paste
  - lexicographic manual can be consulted context sensitive within *Oxygen*

3.1 An example: *elexiko*

- Workflow:
  - *Oxygen* validates entries, when checking the entry into the *ORACLE*-database, a second validation takes place
  - *Vernetziko* is used extensively to create links and to maintain them
  - only after extensive proofreading and testing (in an in-house browser preview), the entry is published online

3.1 A critical reflection on the *elexiko* environment

- *Oxygen* offers many of the funtionalities of DWSs, e.g. copy & paste of parts of entries, validation, WYSIWYG- and 'tree-diagram' view
- *ORACLE*-database supports server-client-architecture, enables data import and export and offers sophisticated search options
- Own tools support workflow or linking entries; but: support of workflow is not very elaborate

3.1 A critical reflection on the *elexiko* environment

- Investment in the beginning of the project for *ORACLE* was rather high; thus, no off-the-shelf DWS was bought, but with little technical effort, writing the entries could start
- In the meantime, total expenditure for all the tools was definitely higher than cost for purchase and support for an off-the-shelf DWS would have been
- But: since all dictionaries in the IDS dictionary portal use the same technical environment, cost is justifiable

3.1 A critical reflection on the *elexiko* environment

- Hosting, maintenance, back-up, and training for our environment is in-house
- The environment is compatible to *Windows*-PCs
- But: Several different projects and persons are in charge of the different components; thus, with technical problems, it is not always easy for lexicographers to know who is responsible

4. Conlusions & outlook

- Some suggestions:
  - in the phase of planning, a thorough analysis of off-the-shelf-DWSs is mandatory
  - own developments should offer the same features as any commercial DWS
  - own developments can be useful to the whole lexicographic community if they are developed as share ware
  - when training lexicographers, detailed information on DWSs should not be forgotten

Thank you for your attention!




Some references

Dictionaries

elexiko. 23.07.2012. In: Institut für Deutsche Sprache (ed.). *OWID - Online Wortschatz-Informationssystem Deutsch*. Mannheim. http://www.elexiko.de.

OWID – Online Wortschatz-Informationssystem Deutsch. 23.07.2012. Institut für Deutsche Sprache (ed.). Mannheim. http://www.owid.de.

Dictionary Writing Systems

ABBY Lingvo Content. 23.07.2012. http://www.abbyy.com.

DEB. 23.07.2012. http://deb.fi.muni.cz/index.php.

IDM DPS. 23.07.2012. http://www.idm.fr.

iLEX. 23.07.2012. http://www.emp.dk.

Termania Portal. 23.07.2012. http://www.termania.net/.

TLex. 23.07.2012. http://tshwanedje.com.

Bibilography

Abel, A. im Druck. 'Dictionary Writing Systems and Beyond.' In: S. Granger & M. Paquot (eds.), *Electronic Lexicography*. Oxford: Oxford University Press.

de Schryver, G.-M. 2011. 'Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters.' In: K. Akasu und S. Uchida (eds.), *ASIALEX 2011 Proceedings*.

Kilgarriff, A. 2006: Word from the Chair. In: de Schryver, G. M. (ed.): *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems*, 7. Pretoria: $(SF)^2$ Press.

Meyer, P. 2011. '*vernetziko*: A Cross-Reference Management Tool for the Lexicographer's Workbench.' In: I. Kosem &K. Kosem (eds.), *Electronic lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex2011, Bled, Slowenien, 10 - 12 November 2011.* Ljubljana: Trojina, Institute for Applied Slovene Studies, 191-198. (http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-25.pdf).

Rundell, M. und A. Kilgarriff 2011. 'Automating the creation of dictionaries: where will it all end?' In: F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (eds.): *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: John Benjamins Publishing Company, 257-282.