Integrating lexical resources through an aligned lemma list

Axel Herold Lothar Lemnitzer Alexander Geyken

Berlin-Brandenburgische Akademie der Wissenschaften

March 9, 2012 @ LDL-2012

Project background

DWDS

- “Digitales Wörterbuch der deutschen Sprache”
  `http://www.dwds.de/`
- lexical information system (Klein, 2004):
  - corpora
  - lexical resources
  - statistical views

CLARIN-D

- “Common language resources and technology infrastructure”
  `http://de.clarin.eu/`
- research infrastructure for language data:
  - interoperability
  - persistence

Linking across dictionaries I

General idea

- goal: confirmed *explicit* linking across lexical resource
- what resolution (i. e. which elements should be linked: lemmas, entries, senses, . . .)?
- which relations?
- aplication: information aggregation, overlap between resources
- broad audience

⟶ we will link "semantically equivalent" entries

Linking across dictionaries II

Terminology

entry *E* basic item of a dictionary,
comprises lexical information for a *lemma*

lemma *L* abstract linguistic sign
(e.g. for a concept expressable as morpheme, word, phrase, . . .); represented by *headwords*

headword *H* selected instance from the lemma's morphological paradigm ("canonical form")

sense *S* semantic space

$$E := \{L_1, \ldots, L_m\} \cup \{S_1, \ldots, S_n\}$$

$$L := \{H_1, \ldots, H_m\}$$

$\longrightarrow$ heuristic: headword similarity as estimator for entry equivalence

Linking across dictionaries III

The dictionaries

eWDG2 model of (WDG, 1962–1977), a 6 volume print dictionary; TEI P5 markup; 120 000 full entries + 40 000 attested lemmas

DWDSWB based on eWDG2; new and continously extended edition of (WDG, 1962–1977); 25 000 additional entries projected for the next six years

EtymWB model of (Pfeifer, 1989), a 3 volume print dictionary; TEI P5 markup; 8 000 morphologically simplex lemmas + 14 000 derivations (and few compounds)

^{1}DWB model of (DWB, 1854-1961), a 33 volume print dictionary; TEI P5 markup (largely typographic view); $300\,000 + X$ main entries, $22\,000 + Y$ related entries

⟶ vastly diverse information, partly only shallow structuring

Linking across dictionaries IV

Entry equivalence

- “entries that describe the same lemma”
- problematic because lemmatization is theory dependent:
    - homonymy, arbitrary ordering of homonyms
    - incompatibility: different numbers of homonyms acknowledged and/or selected
    - headword variance: different canonicalization strategies; historical orthography
    - ideosyncratic headwords: capitalization and character substitution in ^{1}DWB
    - historical and dialectal lemmas in ^{1}DWB; “neologisms” in other dictionaries
    - semantic change: meaning shift, substitution

Linking across dictionaries V

Headword related problems

- unregulated orthography in ^{1}DWB, older orthographic norms
- ideosyncratic canonical headword forms
- regular choice among canonical headword forms

Examples

$^{\text{WDG}}$**Tür** $\equiv$ $^{1\text{DWB}}$**THÜR, THÜRE** ‘door’

$^{\text{WDG}}$**Spaß** $\equiv$ $^{1\text{DWB}}$**SPASZ** ‘fun, joke’

$^{\text{WDG}}$**Beamte** [m.] $\equiv^{\text{DWDSWB}}$**Beamte** [m. f.]
$\equiv^{1\text{DWB}}$**BEAMTE** [m.]
$\equiv^{\text{EtymWB}}$**Beamter** [m.] ‘civil servant’

$^{\text{WDG}}$**Aliment** [n.] $\equiv^{1\text{DWB}}\emptyset$ $\equiv^{\text{EtymWB}}$**Alimente** [plur.] ‘alimony’

Linking across dictionaries VI

Homonymy related problems

homography, actually (Behrens, 2002)

$H_1 = H_2$, but

- different grammatical features in L_1 and L_2
  (formal criterion)
- only unrelated senses
  (semantic/etymological criterion)

Examples

$^{\text{WDG}}$**See**, m. $\equiv$ $^{1\text{DWB}}$**SEE**, m. f. $\equiv$ $^{\text{EtymWB}}$**See**, m. f. (‘lake’)
$^{\text{WDG}}$**See**, f. $\equiv$ $^{1\text{DWB}}$**SEE**, m. f. $\equiv$ $^{\text{EtymWB}}$**See**, m. f. (‘ocean’)

$^{\text{WDG}}\emptyset \equiv$ $^{1\text{DWB}}$**ART^2**, f. $\equiv$ $^{\text{EtymWB}}$**Art^1**, f. m. (‘ploughed land’)
$^{\text{WDG}}$**Art**, f. $\equiv$ $^{1\text{DWB}}$**ART^1**, f. $\equiv$ $^{\text{EtymWB}}$**Art^2**, f. (‘nature, type’)

Linking across dictionaries VII

Sense related problems

- different number of senses (synchronic/separating vs. diachronic/integrating presentation)
- semantic shift

Examples

$^{\text{WDG}}$**gebildet**		$^{1\text{DWB}}$**GEBILDET**	
sense 1	$\equiv$	sense 3	(‘educated, intellectual’)
[derivational base]	$\equiv$	sense 2	(‘shaped, made of’)
$\emptyset$	$\equiv$	sense 1	(‘illustrated’)
$^{\text{WDG}}$Trillion	$\not\equiv$	$^{1\text{DWB}}$TRILLION	(10^{18} vs. 10^{15})

(also very common for artifacts)

Linking across dictionaries VIII

Equivalence across WDG and ^{1}DWB

- ≈ 45 000 common headword forms
- random sample of 941 entries
- 67 % (632) clearly equivalent
- 3 % (27) clearly not equivalent
- 8 % (79) overlap or intersect
- 22 % (203) undecideable (no semantic description)

Open issues

- more relaxed headword mapping via CAB (Jurish, 2010)
- elobarate equivalence categories
- (resolve yet undecidable cases)

Conclusions

Summary

- considerable semantic change over 100 years
- linking entries still requires analysis at sense level
- entry equivalence not a binary relation

Representation

- mapping in RDF or LMF
- persistent identifiers for entries
- dictionaries are *not* provided as triples

Future work

- exploiting the mapping on `http://www.dwds.de/`
- exploiting the mapping in CLARIN-D
- add mapping to GermaNet

Thank you!

Bibliography

Behrens, L. (2002). Structuring of word meaning II: Aspects of polysemy. In D. A. Cruse, F. Hundsnurscher, M. Job, & P. R. Lutzeier (Eds.), *Lexikologie -– Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen* (Vol. 1, p. 319-337). Berlin: de Gruyter.

DWB. (1854-1961). *Deutsches Wörterbuch*. Leipzig: Hirzel.

Jurish, B. (2010). More than words. Using token context to improve canonicalization of historical German. *JLCL*, *25*(1), 23-40.

Klein, W. (2004). Vom Wörterbuch zum Digitalen Lexikalischen System. *Zeitschrift für Literaturwissenschaft und Linguistik*, *136*, 10-55.

Pfeifer, W. (1989). *Etymologisches Wörterbuch des Deutschen*. Berlin: Akademie-Verlag.

WDG. (1962–1977). *Wörterbuch der deutschen Gegenwartssprache*. Berlin: Akademie-Verlag.