Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries

Ina Rösiger, Johannes Schäfer, Tanja George, Simon Tannert,
Ulrich Heid, Michael Dorna

Universität Stuttgart, Universität Hildesheim, Robert Bosch GmbH

elex-2015: Herstmonceux, August 2015

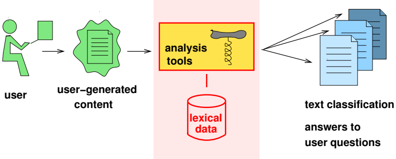
Overview

- Context and objectives
- Data and technology:
  Corpus linguistic tools: components and evaluation
- Extraction of relational data from texts:
  taxonomic and non-taxonomic relations between domain objects
- Sample results
- Collecting data for lexicographic purposes
- Conclusion and future work

Objectives

General context

- Data on the internet:
  - Domain-specific user-generated content: UGC
    forums, discussion groups, etc.,
    from the field of do-it-yourself instructions.
  - Expert-produced texts from the same domain: EXP
    manuals, handbooks, articles, ...

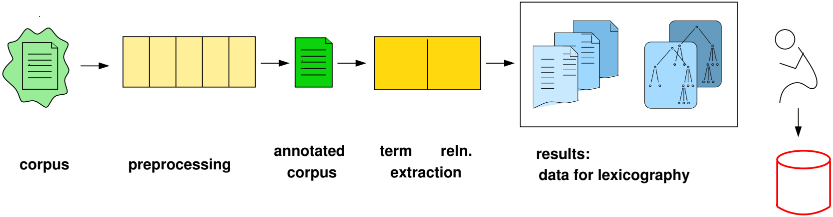


- Need for professional text analysis:
  - Tools to analyze the UGC
    from a domain-related viewpoint:
    classification by topics,
    finding answers for (e.g. forum) questions, etc.
  - **Lexical resources to feed the tools:**
    * To be created interactively
    * To be used both interactively and/or automatically

Lexicographic objectives

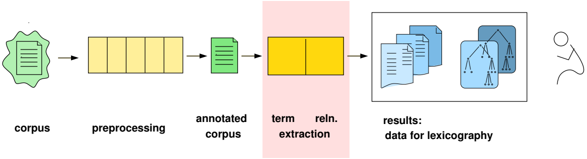
Identifying raw material for interactive e-dictionary building

- Scenario:
  - Automatic extraction of candidate data from corpora
    to create entries of a specialized dictionary
    * term candidates
    * term variants – phraseological variants
    * taxonomic and non-taxonomic relations, e.g. "made-of", "serves-for"...
  - Collecting data for interactive entry construction



Lexicographic objectives

Focus in this presentation



- Not on dictionary as an end product
- But on tools for
  - term candidate extraction
  - extraction of relational data
- Why not use tools like the *SketchEngine*? Kilgarriff et al. 2004
  - Relation extraction requires specific procedures
  - Specialized corpora are small: issue for statistical tools 2.7 - 17.9 M
  - Requirements of work on German data:
    - Dependency parsing
    - Analysis of German compounds

Case study: Data used

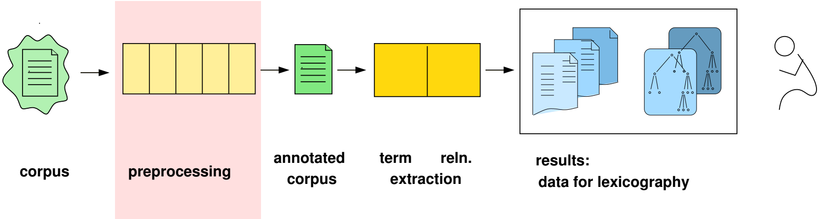
Texts from the do-it-yourself domain (DIY)

- So far: opportunistic collection
- Different genres, text types, etc.
- EXP:UGC = 1:4.3
- Subset used in evaluation: 2.7 M

Text types	aut.	size (w)	totals
DIY manuals, tool manuals	EXP	131,254	
DIY (web) encyclopedias	EXP	28,430	
Tool test reports	EXP	239,238	
Marketing texts	EXP	35,302	
DIY articles, "tricks",etc.	EXP	2,807,487	
Total: expert texts			3,241,711
DIY project descriptions	UGC	4,479,437	
DIY forum posts	UGC	7,873,115	
Forum FAQs, articles, etc.	UGC	450,143	
Wiki content	UGC	896,267	
Total: user-generated texts			13,698,962
varia (without metadata)	?	961,236	
total: data collection			**17,901,909**

Technology used

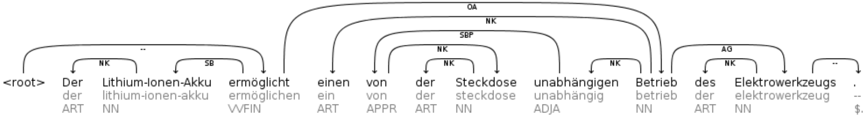
Term candidate extraction – overview of preprocessing steps



- Standard corpus technology for preprocessing
  - Tokenizing [Schmid 2000]
  - Tagging, Lemmatization: RF-Tagger [Schmid/Laws 2008]
  - Dependency parsing: *mate* [Bohnet 2010, Björkelund et al. 2010]

Technology used

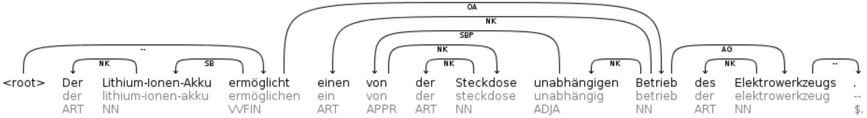
Term candidate extraction – preprocessing: parsed data



- Der Lithium-Ionen-Akku ermöglicht einen von der Steckdose unabhängigen Betrieb des Elektrowerkzeugs
  “The Lithium ion accumulator enables an operation of the power tool which is independent from the socket”

Technology used

Term candidate extraction – preprocessing: parsed data

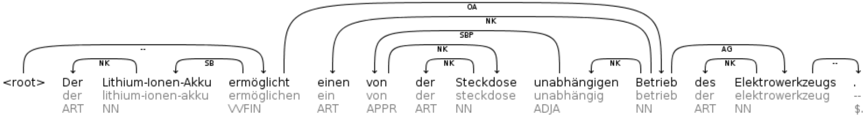


0	**Der**	**SUBJ-Embedded**	The
1	**Lithium-Ionen-Akku**	**SUBJ-Head**	lithium ion accumulator
2	**ermöglicht**	**VERB-Active**	enables
3	**einen**	**OBJ-Embedded**	a
4	**von**	**OBJ-Embedded**	from
5	**der**	**OBJ-Embedded**	the
6	**Steckdose**	**OBJ-Embedded**	socket
7	**unabhängigen**	**OBJ-Embedded**	independent
8	**Betrieb**	**OBJ-Head**	operation
9	**des**	**OBJ-Embedded**	of the
10	**Elektrowerkzeugs**	**OBJ-Embedded**	power tool
11	.	NULL	.

- Standard dependency representation:
  - **verb**
  - **subject**
  - **object**

Technology used

Term candidate extraction – preprocessing: parsed data

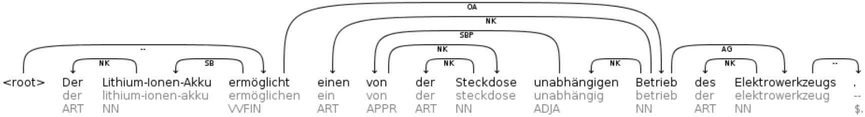


0	Der	SUBJ-Embedded	The
1	Lithium-Ionen-Akku	SUBJ-Head	lithium ion accumulator
2	ermöglicht	VERB-Active	enables
3	einen	OBJ-Embedded	a
4	von	OBJ-Embedded	from
5	der	OBJ-Embedded	the
6	Steckdose	OBJ-Embedded	socket
7	unabhängigen	OBJ-Embedded	independent
8	Betrieb	OBJ-Head	operation
9	des	OBJ-Embedded	of the
10	Elektrowerkzeugs	OBJ-Embedded	power tool
11	.	NULL	.

- Additional tool: extraction from different levels of annotation
  - **heads** of subjects and complements
  - **embedded elements** of subjects and complements
  - **adjuncts** – not part of subjects or complements

Technology used

Term candidate extraction – preprocessing: parsed data

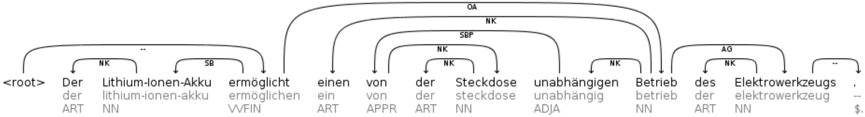


0	Der	SUBJ-Embedded	The
1	Lithium-Ionen-Akku	SUBJ-Head	lithium ion accumulator
2	**ermöglicht**	**VERB-Active**	enables
3	einen	OBJ-Embedded	a
4	von	OBJ-Embedded	from
5	der	OBJ-Embedded	the
6	Steckdose	OBJ-Embedded	socket
7	unabhängigen	OBJ-Embedded	independent
8	Betrieb	OBJ-Head	operation
9	des	OBJ-Embedded	of the
10	Elektrowerkzeugs	OBJ-Embedded	power tool
11	.	NULL	.

- Extraction from different levels of annotation:
  - **heads** of subjects and complements
  - **embedded elements** of subjects and complements

Technology used

Term candidate extraction – preprocessing: parsed data

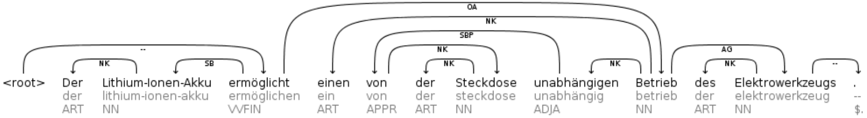


0	Der	SUBJ-Embedded	The
1	**Lithium-Ionen-Akku**	**SUBJ-Head**	lithium ion accumulator
2	**ermöglicht**	**VERB-Active**	enables
3	einen	OBJ-Embedded	a
4	von	OBJ-Embedded	from
5	der	OBJ-Embedded	the
6	Steckdose	OBJ-Embedded	socket
7	unabhängigen	OBJ-Embedded	independent
8	**Betrieb**	**OBJ-Head**	operation
9	des	OBJ-Embedded	of the
10	Elektrowerkzeugs	OBJ-Embedded	power tool
11	.	NULL	.

- Extraction from different levels of annotation:
  - **heads** of subjects and complements
  - **embedded elements** of subjects and complements

Technology used

Term candidate extraction – preprocessing: parsed data



0	**Der**	**SUBJ-Embedded**	The
1	**Lithium-Ionen-Akku**	**SUBJ-Head**	lithium ion accumulator
2	**ermöglicht**	**VERB-Active**	enables
3	**einen**	**OBJ-Embedded**	a
4	**von**	**OBJ-Embedded**	from
5	**der**	**OBJ-Embedded**	the
6	**Steckdose**	**OBJ-Embedded**	socket
7	**unabhängigen**	**OBJ-Embedded**	independent
8	**Betrieb**	**OBJ-Head**	operation
9	**des**	**OBJ-Embedded**	of the
10	**Elektrowerkzeugs**	**OBJ-Embedded**	power tool
11	.	NULL	.

- Extraction from different levels of annotation:
  - **heads** vs. **embedded elements** of subjects and complements

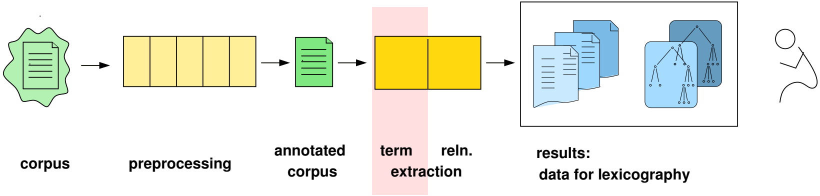
⟹ Information about both:
Grammatical function and span of sentence constituents

combined advantages of dependency and constituency

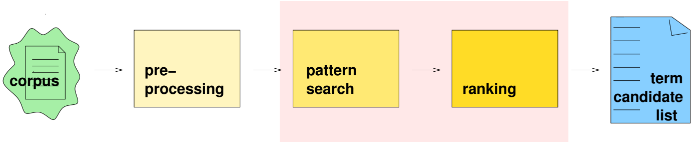
Technology used

Term candidate extraction – patterns and simple statistics

- General overview

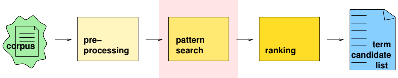


- Term extraction procedures – two steps:
  extraction by patterns – ranking of extracted candidates



Technology used

Term candidate extraction via patterns



- Pattern-based search:
  1. POS-shapes:

- N	Bohrmaschine	"drill"
- Adj + N	oszillierende Säge	"oscillating drill"

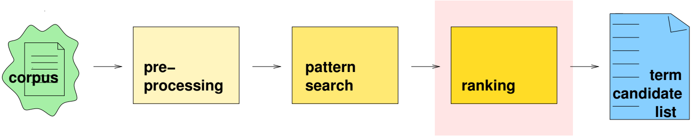
  2. Term-relevant structures extracted from dependency parses:

- N + PP	Bohrer mit Kabel	"drill with cord"
- V + NP_{object}	Temperatur + erhöhen	"increase + temperature"

- Relating simple patterns with more complex patterns to find term variants and their relationship with basic terms:
  3. Subtype-denoting: e.g. ((Adv)? (Adj)? **Adj**)? **N**:
     Farbe → *weiße Farbe* "colour → white colour"
  4. Patterns finding cases of embedded term use:
     (**N Det**)? ((Adv)? Adj)? ((Adv)? **Adj**)? **N**
     bodengleiche Dusche → *Aufbau einer bodengleichen Dusche*
     "walk-in shower → installation of a walk-in shower"

Technology used

Statistical term candidate ranking



- Ranking according to statistical measures: comparison between general-language and domain-specific candidate frequencies
- Domain corpus: DIY data
- General-language corpus: SDeWaC (880 M. tokens)
- Test of several termhood measures Schäfer et al. submitted
- In current experiments: domain specificity Ahmad et al. 1992

Output of term extraction: evaluation

Gold standard-based evaluation

- Gold standard (gs) George 2014
  - 2.7 M sample from the DIY corpus
  - 3 independent annotators
  - basic patterns only: `N`, `Adj+N`, `N+N`$_{Gen}$, `N+Prp+N`
  - Decision: [+/– terminologically relevant]
  - We keep track of {3:0}-decisions (strict)
    and of {2:1}-decisions (liberal)
- Evaluation experiments:
  - Our tool (basic version) ↔ SDL (Trados) Multiterm Extract
  - Different termhood measures Schäfer 2015
  - Use of additional (dependency-) syntactic filters

Schäfer et al. submitted

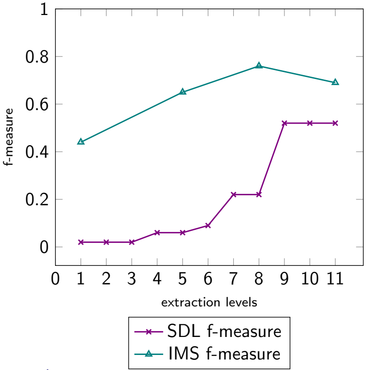
Output of term extraction: examples from the evaluation

Quantitative results – comparison with SDL MultiTerm Extract

- Best f-measures per tool

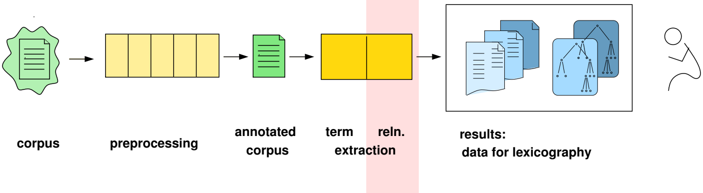
		liberal gold standard				
pattern:		N+“von”+N	N+N_{gen}	N	ADJ+N	N+Prp+N
IMS	Precision	72%	65%	52%	38%	55%
	Recall	84%	91%	85%	55%	73%
	F-measure	0.78	0.76	0.65	0.45	0.63
SDL	Precision	66%	40%	39%	33%	44%
	Recall	68%	76%	76%	22%	73%
	F-measure	0.67	0.52	0.52	0.26	0.55

- F-measure in terms of quality levels of SDL's tool: **Noun + Noun**$_{Genitive}$



Extracting data on relations

Overview



- Taxonomic relations:
  - Building partial hierarchies
    of superordinate and subordinate domain objects:
    by means of taxonomy patterns and compound analysis
  - Including term variants
- Non-taxonomic relations:
  Collecting data by means of
  an analysis of compounds and their syntactic variants

Extracting data on taxonomic relations

Combining different methods

- Taxonomy patterns: cf. Hearst 1992 etc.
  - *an X is a Y which...*
  - *X_1, X_2,... and other Ys...*

  ⇒ relevant for relations between items that are not morphologically related

- Analysing German compounds:
  `X·Y` is a type of `Y`:
  - *Band·säge* → *Säge* band·saw → saw
  - *Mehrzweck·werkbank* → *Werkbank* multi-purpose·workbench → workbench

- Syntactic analysis of phrases expressing taxonomic relations:
  - `Adj+N` is a type of `N`:
    durchsichtige Farbe → *Farbe* transparent colour → colour
  - `Adv + Adj + N` is a type of `Adj+N`:
    matt weiße Farbe → *weiße Farbe* matt white colour → white colour

Extracting data on taxonomic relations

Analysis of German compounds – methodology

- Compound splitting with COMPOST, Cap 2014
  a hybrid tool based on morphological rules and corpus data
  - Head as superordinate
  - Compounds considered as subtypes of their heads:
    Säge → {*Kreissäge, Bandsäge, ...*} saw → circular saw, bandsaw ...

- Implementation is aware of complex non-heads:
  (1) Split into morphemes:
  Eigenbaubandsäge → *eigen* · *bau* · *band* · *säge*
  self-made · bandsaw → *self* · *build* · *band* · *saw*
  (2) Check for attested morpheme combinations:
  - Bandsäge bandsaw
  - *Baubandsäge *construction bandsaw
  - Eigenbau-X: *Eigenbaumöbel, Eigenbauschlitten, etc.*
    self-made furniture, self-made sledge, etc.

  (3) Correct split: *Eigenbau·Bandsäge*

Extracting data on taxonomic relations

Analysis of German compounds – sample results

Candidate		**Analysis**
Bandsäge	bandsaw	Band|Säge
Elektro-Bandsäge	electric bandsaw	Elektro|Band|Säge
Hand-Bandsäge	hand bandsaw	Hand|Band|Säge
Horizontalbandsäge	horizontal bandsaw	Horizontal|Band|Säge
Vertikalbandsäge	vertical bandsaw	Vertikal|Band|Säge
Metallbandsäge	metal bandsaw	Metall|Band|Säge
Minibandsäge	mini bandsaw	Mini|Band|Säge

Extracting data on taxonomic relations

Sample results: Combining patterns and compound analysis

← Hypernyms - Hyponyms →	Tools used	Gloss
Elektrowerkzeug		power tool
- Schleifer	taxonomic pattern	sander
- Bandschleifer	compound analysis	belt sander
- Exzenterschleifer	compound analysis	random orbital sander
Elektrowerkzeug		power tool
- Kreissäge	taxonomic pattern	circular saw
- Handkreissäge	compound analysis	circular handsaw
- Tischkreissäge	compound analysis	circular table saw

Extracting data on non-taxonomic relations

Combining compound splitting and the search for syntactic paraphrases

- Compound splitting using COMPOST Cap 2014
- Use of head and non-head items in pattern search:
  different syntactic patterns, depending on type of the head
  - nominal heads:
    - * $N_1 \cdot N_2 \longrightarrow N_2 + \text{Prep} + N_1$:
      Schraubenloch ⟶ *Loch für Schraube* 'screw·hole' → hole for screw
    - * $N_1 \cdot N_2 \longrightarrow N_2 + N_{1-\textit{Genitive}}$:
      Raummitte ⟶ *Mitte des Raums* 'room·centre' → centre of room
  - deverbal heads:
    - * $N_1 \cdot V_2^n \longrightarrow V_2^n + N_{1-\textit{Genitive}}$:
      Temperaturerhöhung ⟶ *Erhöhung der Temperatur*
      'temperature·increase' → increase of temperature
    - * $N_1 \cdot V_2^n \longrightarrow V_2 + \text{Obj}(N_1)$:
      Holzbohrer ⟶ *Holz + bohren [jmd. bohrt Holz]*
      'wood·drill' → (to) drill + wood [sbdy drills wood]

Extracting data on non-taxonomic relations

Purpose and sample results

(1) Getting more evidence for a term candidate:

		f_{cmpd}	f_{synt}	$\sum$
– *Schraubenloch* (screw+hole)	↔ *Loch für Schraube* (hole for screw)	441	15	456
– *Raummitte* (room+center)	↔ *Mitte des Raumes* (center of the room)	37	57	94
– *Holzmaserung* (wood+grain)	↔ *Maserung des Holzes* (grain of the wood)	136	56	192
– *Brettkante* (board+edge)	↔ *Kante des Brettes* (edge of the board)	79	41	120

(2) Data for specific types of relations:

material:		preposition: *aus* (made of)	
	Stahlschraube	↔ *Schraube aus Stahl*	(steel screw)
	Edelstahlschraube	↔ *Schraube aus Edelstahl*	(stainless steel screw)
	Kupferschraube	↔ *Schraube aus Kupfer*	(copper screw)
application:		preposition: *für* (for)	
	Rigips-Schraube	↔ *Schraube für Rigips*	(screw for plasterboard)
property:		preposition: *mit* (with)	
	Senkkopf-Schraube	↔ *Schraube mit Senkkopf*	(countersunk head screw)
purpose:		preposition: *als/zu* (as/to)	
	Führungsschraube	↔ *Schraube als Führung*	(screw as a guide)
	Befestigungsschraube	↔ *Schraube zu Befestigung*	(screw as a fixing)

Collecting data for lexicographic work

Principles: implementation is ongoing

- All tool components are applied to the DIY corpus
- Each tool produces
  - result data,
    to be sorted by “central” lexical items: e.g.
    - base of V + N_{OBJ} collocation
    - head of compound
  - For each item of the result data:
    process metadata, to indicate provenience:
    - textual source
    - tool (component) used
- Under way:
  Tool to collect all these data per “central” item
  and to display it

Collecting data for lexicographic work

Example of a (partial) data collection: s.v. *Schraube* (screw)

- Adjectives and related compounds and multiword variants:

– lang – kurz; groß – klein	long – short; big – small
– versenkt, seitlich versenkt;	countersunk, laterally countersunk
Senkkopfschraube	countersunk head screw
– metrisch,	metrical
Schraube mit metrischem Gewinde	screw with metrical thread
– rostfrei, feuerverzinkt	stainless, hot-galvanized

- Multiword terms with PPs:

– Schraube mit Sechskantkopf,	screw with hexagon head
Sechskantschraube	hexagon screw
– Schraube mit zylindrischem Kopf	screw with cylindrical head
– Gewinde der Schraube	thread

- Verbal contexts:

– Schraube$_{Obj}$ (ein)drehen, (ein)schrauben; Eindrehen d. S.	screw (in)
– Schraube$_{Obj}$ anziehen, festziehen	tighten
– Schraube$_{Obj}$ lösen, entfernen	remove, unscrew

Conclusion

- Has been shown:
  - A set of tools to extract terms and their relations
  - Sample results from the DIY domain
  - Proposals for lexicographic use
- Next steps:
  - Enhancement of the tools – more detailed evaluations
  - Implementation of further tools needed for work with UGC: e.g. coreference resolution, to improve extraction of relations involving verbs and their (pronominalized) complements
  - Implementation of tool to combine the output for lexicographic purposes