University of Stuttgart
Germany

What good are ‘Nominalkomposita’ for ‘noun compounds’

Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors

Patrick Ziering Lonneke van der Plas
Institute for NLP, University of Stuttgart, Germany
COLING 2014

1. Controversy of Defining Compoundhood

General existence of compoundhood

There is virtually no reliable/universally accepted definition for compoundhood ([LS09])

- [Bau03]: *formation of a new lexeme by adjoining two or more lexemes*
- [Mar67]: *No compounding word formation:* EXPANSION

Distinction between compounds and phrases

- Is *tomato bowl* a special kind of bowl (i.e., a lexeme)? (cf. *deictic compounds* ([Dow77]))

Solution: Linguistic tests for compoundhood

- Inseparability (*black **ugly** bird*)
- Inability to modify the modifier (***very** social policy*)
- Spelling as one word (*football*, but *waiting room*)

2. First Extraction Iteration

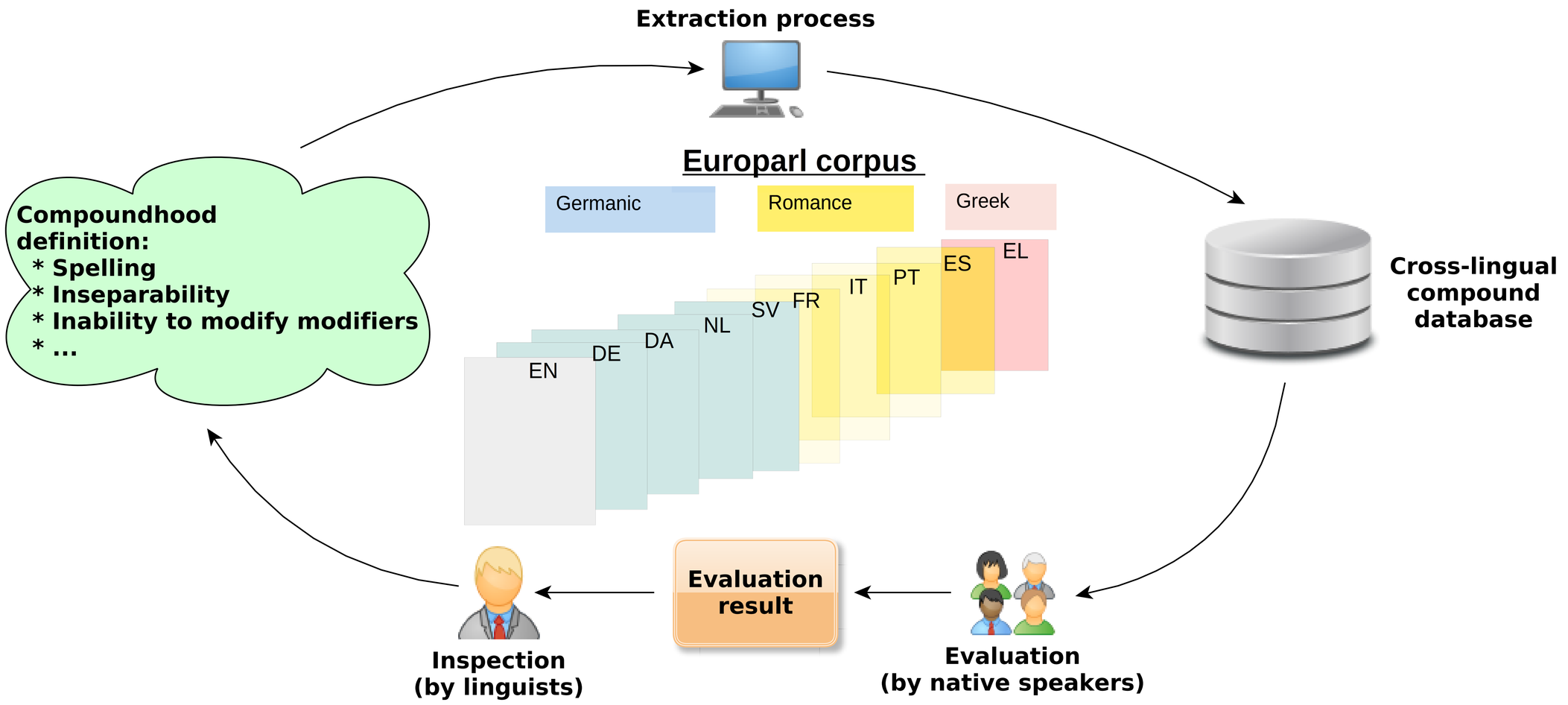
Initial definition

An English word sequence is a compound, if it passes the following linguistic test:

- Spelling as one word - defined cross-lingually

Extraction process

1. Preselection of English compounds using PoS chunks (e.g., noun-prep-noun)
2. PoS error filter (e.g., stop words tagged as noun)
3. Word alignment filter (e.g., clipping determiners)
4. **Closed compound restrictor** for ***n*** languages CCR(n): The English word sequence has to be aligned to a closed compound in at least ***n*** languages (e.g., the German *Wartezimmer* for *waiting room*)
   → Optimal ***n*** for precision/recall trade-off



3. Experiment on First Iteration

Setup

- PoS tagging
- Sentence alignment (884K parallel sentences)
- Word alignment (GIZA++)
- Binary compound splitter (according to [SCA13])

Evaluation

50 accepted/rejected samples for each database
→ Determination of true/false positives/negatives

Database	Size	Precision	Recall	F-Score
Basic database	3,178,661	38.0%	—	—
CCR(1)	795,518	84.0%	71.2%	77.1%
CCR(2)	495,837	92.0%	74.2%	**82.1%**
CCR(3)	316,330	98.0%	65.3%	78.4%
CCR(4)	143,121	98.0%	63.6%	77.2%

Controversial cases

→ further stimulate linguistic discussion:
For example: German A+N compounds (cf. [SH09]):

- *strong wind* ↔ de:*Starkwind*
- *small car* ↔ de:*Kleinwagen*
- *used car* ↔ de:*Gebrauchtwagen*

Collocation or compound?

⇒ Mostly: semantic specification (i.e., a lexeme)

4. Database case study: Bracketing compounds

The task

LEFT or RIGHT branching of tripartite noun compounds (e.g., *[human rights] abuses* or *baby [bicycle seat]*)

Six aligned phrase patterns

For example:

- ADJ CNC
  geplante Bildungsreform ([education reform] plan)
- SN FC CNC
  verslagen over autoprijzen ([car prize] reports)
- SN ADJ FC SN
  consumo final de energía (energy [end consumption])

Method	Accuracy
LEFT baseline	71.1 %
Cross-lingual phrase patterns	**91.6** %

References

[Bau03] Laurie Bauer.
Introducing Linguistic Morphology.
Edinburgh University Press, 2003.

[Dow77] Pamela Downing.
On the creation and use of English compound nouns.
Language, 53(4):810–842, 1977.

[LS09] Rochelle Lieber and Pavol Stekauer.
The Oxford Handbook of Compounding.
Oxford Handbooks in Linguistics. OUP Oxford, 2009.

[Mar67] Hans Marchand.
Expansion, Transposition, and Derivation.
La Linguistique, pages 13–26, 1967.

[SCA13] Sara Stymne, Nicola Cancedda, and Lars Ahrenberg.
Generation of Compound Words in Statistical Machine Translation into Compounding Languages.
Computational Linguistics, 39(4):1067–1108, 2013.

[SH09] Barbara Schlücker and Matthias Hüning.
Compounds and phrases. A functional comparison between German A+N compounds and corresponding phrases.
Italian Journal of Linguistics / Rivista di Linguistica, pages 209–234, 2009.

Supported and funded by the Deutsche Forschungsgemeinschaft (DFG SFB 732-D11)

http://www.ims.uni-stuttgart.de