



Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations

Introduction to the Workshop

Erhard Hinrichs and Verena Henrich

Department of Linguistics and Collaborative Research Center (SFB 833)

University of Tübingen

erhard.hinrichs@uni-tuebingen.de & verena.henrich@uni-tuebingen.de

http://www.sfs.uni-tuebingen.de/~vhenrich/cclcc_2014/



Presentation Outline

- Short personal introduction
- Establishing common ground for terminology
- Motivation for including this workshop in a summer school on logic, language, and information
- Overview of the workshop program
- Some practical matters



Morphologically Complex Words

- Complex words are the result of two types of word formation processes:
 - Derivation
examples: instruct-ion-al, un-happi-ness
 - Compounding
examples: toothbrush, snowman



Multi-Word Expressions (Bauer, 1983; Sag et al., 2002)

- “We define multiword expressions (MWEs) very roughly as ‘idiosyncratic interpretations that cross word boundaries (or spaces)’.” (Sag et al., 2002, p. 2)
- Nonce formation
- Institutionalization
 - Collocations
- Lexicalization
 - Fixed expressions
 - Semi-fixed expressions
 - Syntactically-flexible expressions



Nonce Formation

- New word (neologism) coined in a particular communicative situation with a semantics that is not yet conventionalized
 - Examples: vacation alarm, fish lady



Institutionalization: Collocations

- In the broad sense of the term: any statistically significant co-occurrence of two or more words
 - Examples: smoke a cigarette, buy a newspaper
- In the narrow sense of the term: any statistically significant co-occurrence of two or more words where the meaning of at least one collocate is semantically idiosyncratic
 - Examples: black coffee, heavy smoker, artesian well



Lexicalization: Fixed Expressions (Sag et al., 2002)

- Immutable expressions that defy conventions of grammar and compositional interpretation
 - Examples: ad hoc, Palo Alto



Lexicalization: Semi-Fixed Expressions (Sag et al., 2002)

- Non-decomposable idioms
 - Examples: kick the bucket, shoot the breeze
- Compound nominals
 - Examples: car park, attorney general
- Proper names
 - Examples: San Francisco, Oakland Raiders



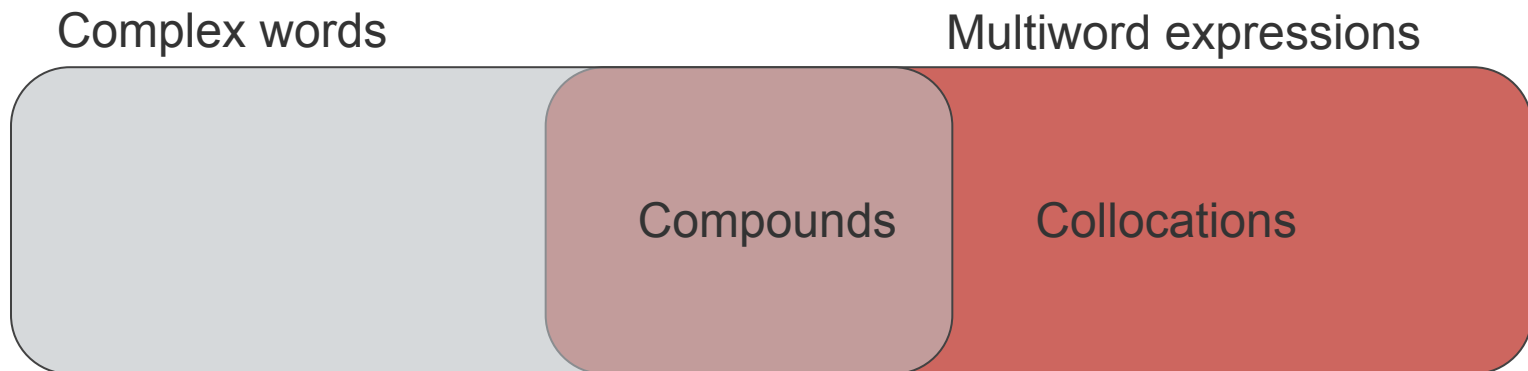
Lexicalization: Syntactically-Flexible Expressions (Sag et al., 2002)

- Verb-particle constructions
 - Examples: look up, brush up on
- Decomposable idioms
 - Examples: spill the beans, sweep under the rug
- Light-verb constructions
 - Examples: make a mistake, give a demo



Back to the Title of Our Workshop

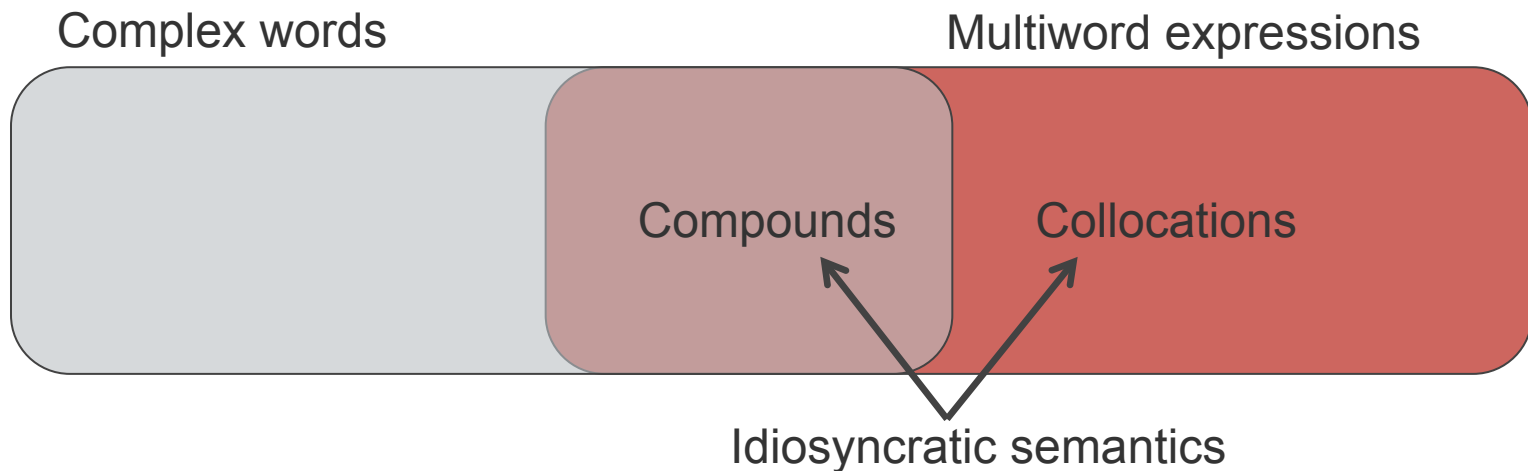
Computational, Cognitive, and Linguistic Approaches
to the Analysis of Complex Words and Collocations





Back to the Title of Our Workshop

Computational, Cognitive, and Linguistic Approaches
to the Analysis of Complex Words and Collocations





CCLCC Workshop

- Question: Why organize a workshop on complex words and collocations as part of a summer school for logic, language, and information?.



- Answer: such expressions raise interesting theoretical, methodological, and practical issues in natural language semantics, in computational linguistics, and in human language processing.



Complex Words and Collocations as a Challenge for Natural Language Semantics

- Principle of compositionality for the interpretation of linguistically complex expressions: the meaning of a complex phrase is a function of the meaning of its parts and the mode of combination of these parts
- Complex words and collocations are units “just above the word level” and thus provide an interesting test bed for the principle of compositionality
- “We define multiword expressions (MWEs) very roughly as ‘idiosyncratic interpretations that cross word boundaries (or spaces)’.” (Sag et al., 2002)



Complex Words and Collocations as a Challenge for Computational Linguistics

- “Multiword expressions are a pain in the neck for NLP”
(Sag et al., 2002)
- MWEs are pervasive in spoken and written language:
 - In WordNet 1.7, 41% of all entries are MWEs
 - In GermaNet 9.0, 62.4% of all nouns are compounds
 - Baroni et al. (2002) report that almost half (47%) of the word types in the APA German news corpus are compounds



Complex Words and Collocations as a Challenge for Human Language Processing

- How are complex words and collocations stored in the brain?
- More generally: are MWEs part of the language faculty or part of general-purpose cognitive structures and memory?
(see Jackendoff (1997, Chapter 7) for a more in-depth discussion)
- How does human processing of complex words compare to human processing of simplex words:
 - In what respects are they similar?
 - In what respects do they differ?



Methodological Issue: How to Model Word Meaning?

- Competing answers:
 - Distributional approaches using word-vector representations
 - Decompositional approaches using logical primitives



Workshop Program

- Contributions:
 - 14 authors from 7 countries
 - Two invited keynote lectures:
 - Dr. Melanie Bell, Anglia Ruskin University, Cambridge, UK
 - Prof. Dr. Eduard Hovy, Carnegie Mellon University, Pittsburgh, PA, USA
- Online on the workshop website:
http://www.sfs.uni-tuebingen.de/~vhenrich/cclcc_2014/program.html



Workshop Program – Day 1

Monday, August 11, 2014

17:00 – 17:30	Erhard Hinrichs and Verena Henrich (University of Tübingen)	Introduction to the Workshop
17:30 – 18:30	Melanie Bell (Anglia Ruskin University)	Invited Talk: Compound stress, informativity and semantic transparency



Workshop Program – Day 2

Tuesday, August 12, 2014		
17:00 – 18:00	Eduard Hovy (Carnegie Mellon University)	Invited Talk: The Semantics of Word Collocations from a Distributional Point of View
18:00 – 18:25	Daniil Sorokin, Corina Dima, and Erhard Hinrichs (University of Tübingen)	Multi-label Classification of Semantic Relations in German Nominal Compounds
18:25 – 18:30	All	Discussion



Workshop Program – Day 3

Wednesday, August 13, 2014 – Computational Methods		
17:00 – 17:25	Pedro Bispo Santos (Technische Universität Darmstadt)	Using compound lists for German decompounding in a back-off scenario
17:25 – 17:50	Eleri Aedmaa (University of Tartu)	Statistical methods for Estonian particle verb extraction from text corpus
17:50 – 18:15	Daria Kormacheva, Lidia Pivovarova, and Mikhail Kopotev (University of Helsinki)	Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams
18:15 – 18:30	All	Discussion



Workshop Program – Day 4

Thursday, August 14, 2014 – Collocations

17:00 – 17:25	Lothar Lemnitzer and Alexander Geyken (Berlin Brandenburgische Akademie der Wissenschaften)	Semantic modeling of collocations for lexicographic purposes
17:25 – 17:50	Zygmunt Vetulani and Grażyna Vetulani (Adam Mickiewicz University)	Verb-Noun Collocations in PoINet 2.0
17:50 – 18:15	Laura Giacomini (University of Heidelberg)	Variational models in collocation: taxonomic relations and collocates inheritance
18:15 – 18:30	All	Discussion



Workshop Program – Day 5

Friday, August 15, 2014 – Multi-Word Expressions

17:00 – 17:25	Konrad Szczesniak (University of Silesia)	Too Colorful To Be Real. The meanings of multi word patterns
17:25 – 17:50	Francesca Quattri (The Hong Kong Polytechnic University)	Cross-language description of shape: Shape-related properties and Artifacts as retrieved from conventional and novel collocations across different languages
17:50 – 18:15	Petya Osenova and Kiril Simov (Bulgarian Academy of Sciences)	Treatment of Multiword Expressions and Compounds in Bulgarian
18:15 – 18:30	Erhard Hinrichs and Verena Henrich (University of Tübingen)	Final Discussion and Closing of the Workshop



Proceedings

- Digital version is available on workshop website:

http://www.sfs.uni-tuebingen.de/~vhenrich/cclcc_2014/CCLCC_2014_workshop_proceedings.pdf

- Printed version
 - Freely available for all participating authors
 - Any additional copies: 5 €



References

Baroni, Marco, Matiasek, Johannes and Trost, Harald (2002). Predicting the components of German nominal compounds. In F. van Harmelen (ed.):, Proceedings of the 15th European conference on artificial intelligence (ECAI), Amsterdam: IOS Press, pp. 470–474.

Bauer, Laurie (1983). English Word-formation, Cambridge: Cambridge University Press.

Jackendoff, Ray (1997). The Architecture of the Language Faculty, Cambridge, MA: MIT Press.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann and Flickinger, Dan (2002). Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh (ed.): Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol. 2276, Springer Berlin Heidelberg, pp. 1–15.