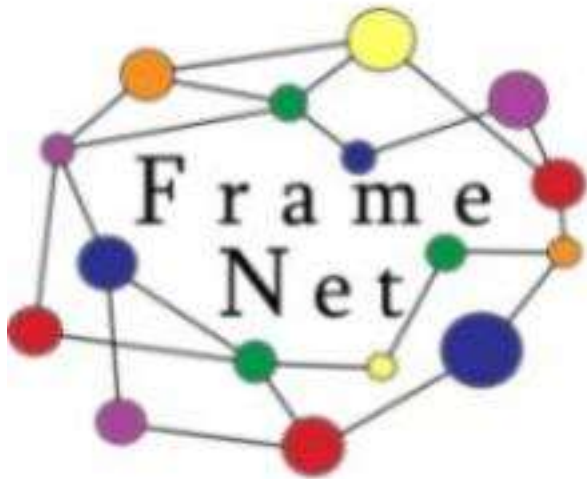


Acquisition of Multiword Lexical Units for FrameNet

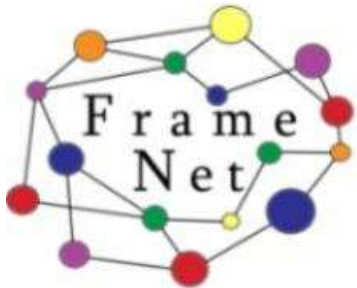
Silvana Hartmann & Iryna Gurevych



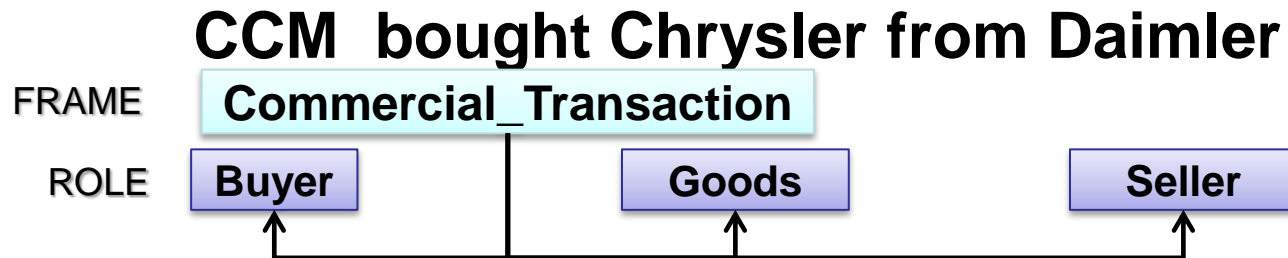
MWE
Multiword Expression

Part 1

MWEs and FrameNet: state of affairs



- Lexical semantic resource modeling „frame semantics“
- **Frames** group words in classes by „typical situations“
- Frame-specific **Roles** are assigned to participants in the situation (syntactic arguments)



MWE

Multiword Expression

- A phrase that stands out because of it is lexicalized or institutionalized (Sag et al., 2002)

- Lexicalized: idiosyncratic syntax, semantics, etc.:

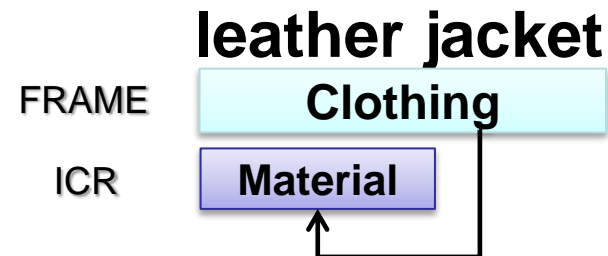
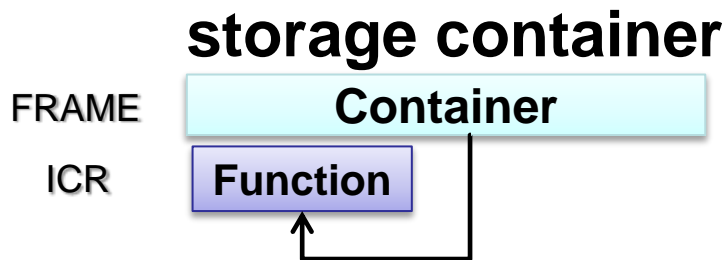
up and away, to spill the beans

- Institutionalized: marked by frequent use:

leather jacket

Modeling MWEs in FrameNet

- Lexical Entries
 - contain MWE constituents („lexemes“)
 - and marks MWE heads
- MWE Semantics:
 - Incorporated Roles (**ICRs**) can be marked
 - Make implicit relations between constituents explicit!
 - See Calzolari et al. (2002)
- Example:



MWEs in FrameNet I

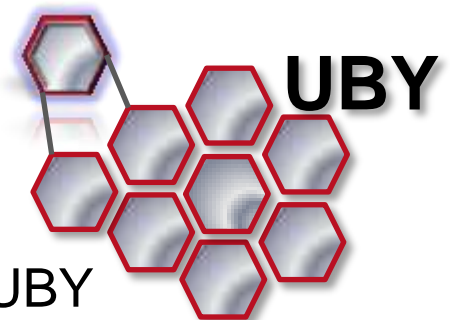
MWE Senses in FrameNet 1.5

POS	# MWEs	Example
Verb	432	sell short, stay in touch
Noun	246	sugar daddy, rush hour
Adjective	125	next generation, heart-warming
Other	63	beyond compare
TOTAL	866	(806 distinct lemmas)

MWEs in FrameNet II

English MWE Lemmas in other resources

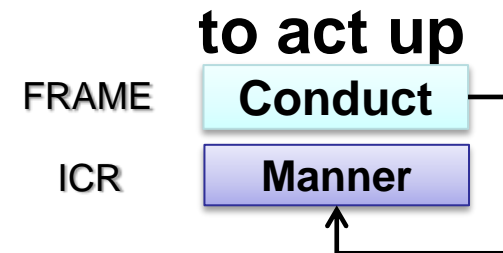
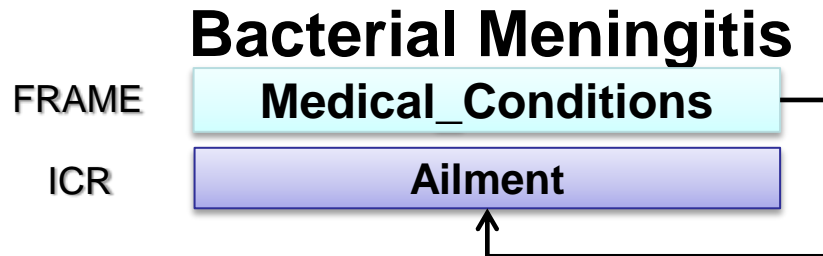
Resource	# MWEs
FrameNet	806
WordNet	69,466
VerbNet	105
Wiktionary EN	67,548
OmegaWiki EN	21,000
ALL	139,281



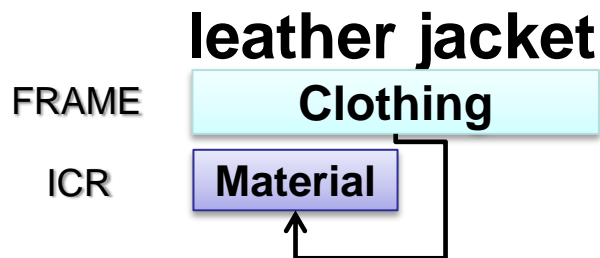
numbers from UBY

MWE Semantics in FrameNet: Incorporated Roles

- Senses with ICRs in FrameNet 1.5: 1145 (56 MWEs)
- Most of the ICRs refer to the full mwe, not a constituent:



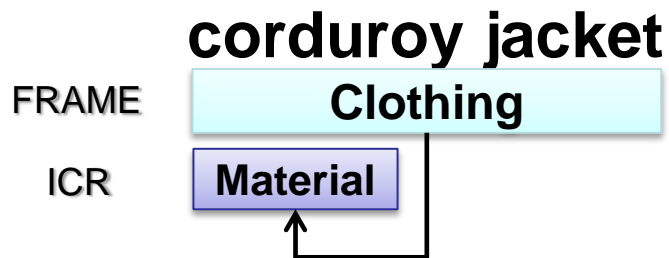
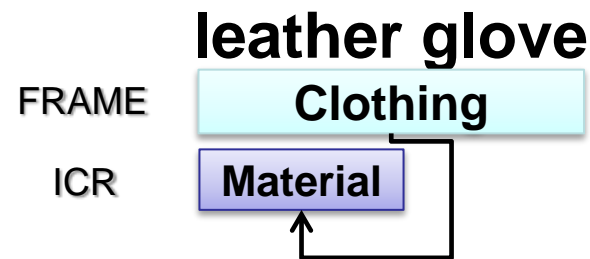
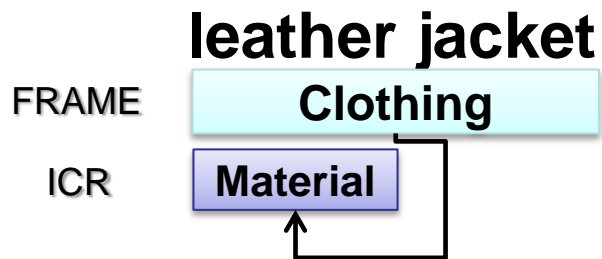
- What is lacking: ICRs referring to constituents



(leather jacket does occur in the example sentences, but not in the lexicon)

MWEs and Constructions

- Productive MWE patterns



- Don't want to list them all
- ➡ derive constructions?

Other Languages

- Example: German

Kalbs|schnittzel
veal / schnitzel

FRAME

Food

ICR

Type
made with



made from veal

Kinder|schnittzel
children / schnitzel

FRAME

Food

ICR

Type
for

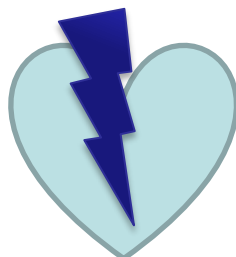
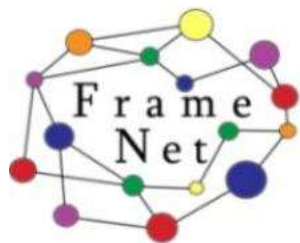


for children

- Also productive: Kinder|döner, Studenten|döner
- Also events: Wetter|vorhersage (*weather|forecast*)

Summary: Situation of MWEs in FrameNet

Currently it doesn't look like a love affair...



MWE
Multiword Expression

SOLUTION

- Add more **(A)** MWEs and **(B)** MWE semantics to FrameNet

GOALS

- Improved representation of MWE semantics and constructions
- Improved Semantic Role Labeling (SRL)
 - Erk & Padó (2006): lack of MWEs harms SRL performance

Acquisition of MWEs and MWE semantics

(A) MWE Acquisition for FrameNet

- Assigning MWEs to Frames
- MWEs from: other lexical semantic resources or corpora
- How?
 - Distributional Methods (see Pennacchiotti et al., 2008)
 - Resource alignments (Johansson & Nuguest, 2007, Laparra & Rigau, 2010, ...)
- Also works for other languages...
 - e.g., Italian (see Pennacchiotti et al., 2008)
 - e.g., alignment to Wikipedia for Italian (Tonelli et al., 2013)
 - e.g., alignment to Wiktionary for German (Hartmann & Gurevych (2013), to appear)



(B) Acquisition of MWE Semantics

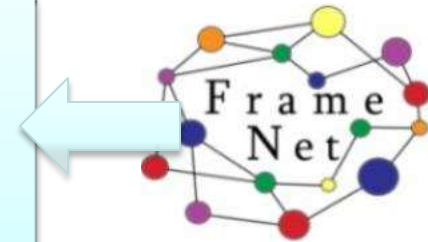
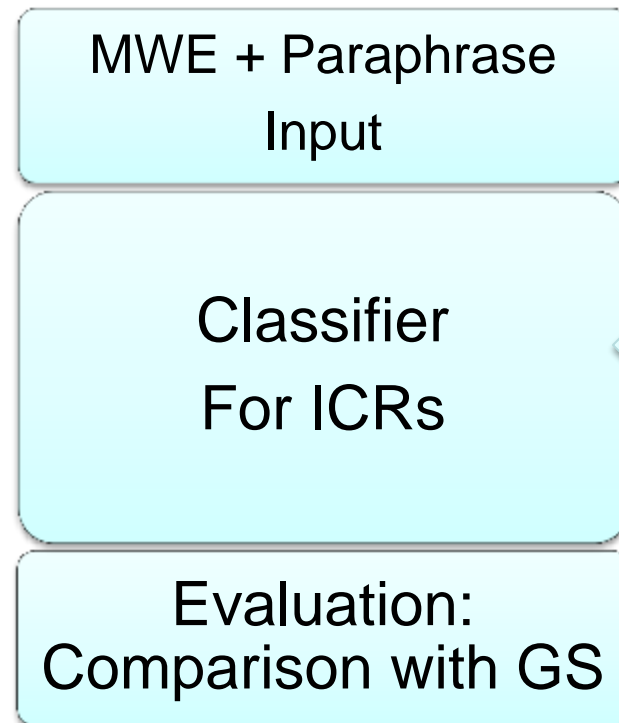
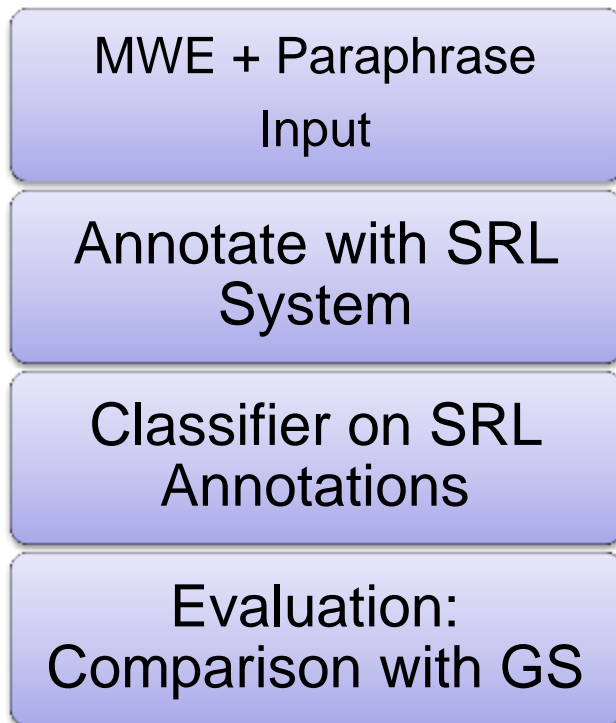
- How to automatically discover ICRs? Use paraphrases!



- Related Task: Noun Compound Interpretation (Nakov, 2008)
- ➡ use datasets from Noun Compound Interpretation

Proposed Setup

- Annotate paraphrases of MWEs with SRL system
- Identify ICRs based on annotations of the paraphrasing words
- Alternative: develop specialized classifier



Proposed Setup

- Annotate paraphrases of MWEs with SRL system
- Identify ICRs based on annotations of the paraphrasing words
- Alternative: develop specialized classifier

MWE + Paraphrase
Input

MWE + Paraphrase
Input

that contains bread

FRAME

Container

ICR

Contents

Evaluation:
Comparison with GS

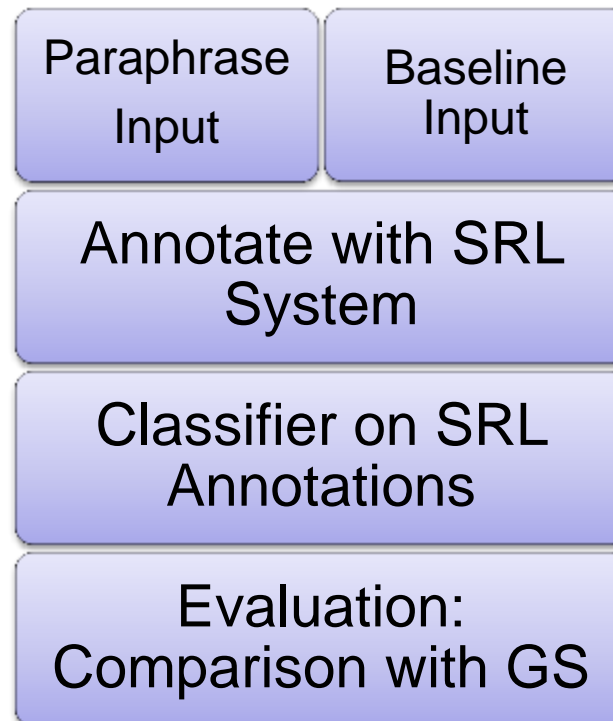
Evaluation:
Comparison with GS

Part 3

First experiments on the acquisition of MWE semantics

First Experiments

- Annotate paraphrases of MWEs with SRL system (SEMAFOR, Das et al., 2010)
- Identify ICRs based on annotations of the paraphrasing words

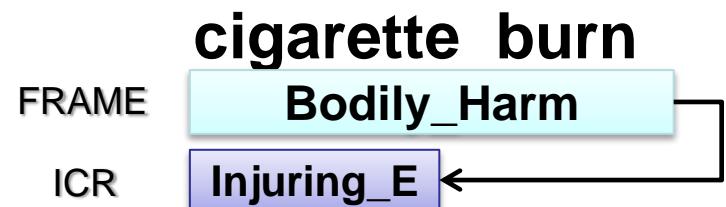


- Noun Compounds and Paraphrases from Semeval Shared tasks:
 - Semeval 2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions (Butnariu et al., 2010)
 - Semeval 2013 task 4: Free Paraphrases of Noun Compounds
- Compounds:
 - Noun compounds with two constituents: adventure story, advertising agency..
 - 812 noun compounds
- Paraphrases:
 - Collected via crowd-sourcing, ranked by frequency
 - Example:

motor yacht  **yacht propelled by a motor**
yacht with a motor

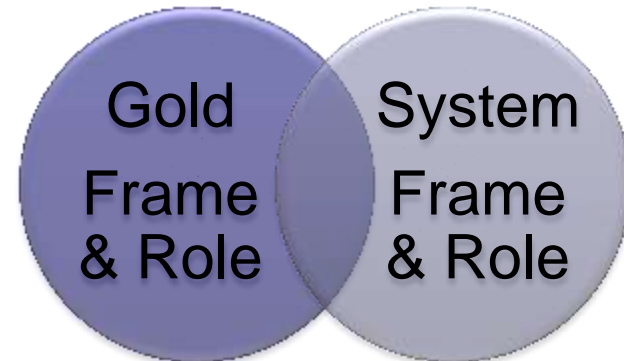
Gold Standard

- **Problem 1:** Minimal overlap with shared task MWEs and FrameNet lexicon: „child abuse“
- **Problem 2:** Few MWEs with ICRs in FrameNet (and no paraphrases for these..)
- **Solution:** annotation of a gold standard (GS) on shared task MWEs
 - Match MWE to candidate frame based on head
 - Two-class annotation: MWE matches the frame (yes|no)
 - If yes: annotate whether non-head matches a role of the frame (= is an ICR)
- 412 MWEs with between 10 and 170 paraphrases each
- 225 have an ICR in the gold standard



Baseline Experiments

- Question: How well does Semafor annotate the MWEs out of context?
- Experiment Setup
 - Annotate MWEs with Semafor
 - Compare System Annotations S to Gold Standard Annotations G
 - High-recall setup:
 - Intersect G frame with S frames
 - If yes, intersect G roles with S roles
 - If set not empty \Rightarrow Role-match
- Results
 - There are annotations for 50% of the GS
 - There is a Frame-match for 125 MWEs
 - There is a Role-match for 10 MWEs

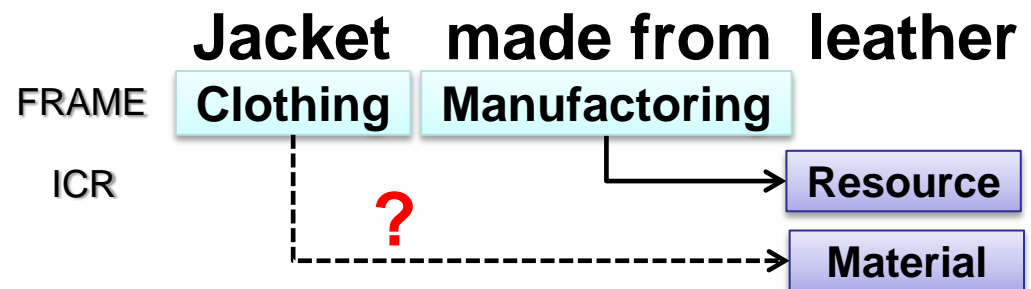
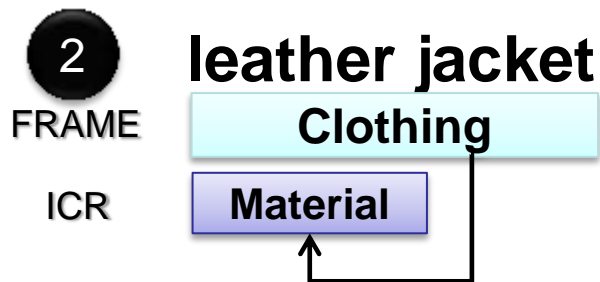
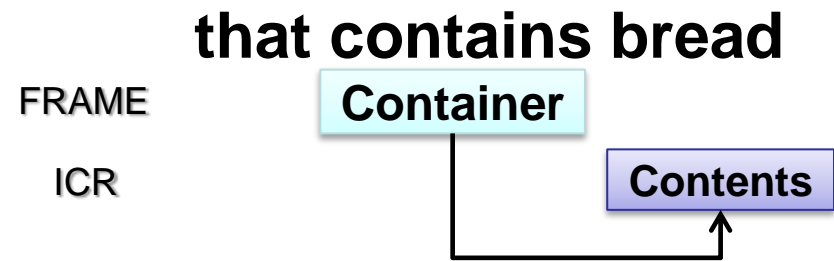
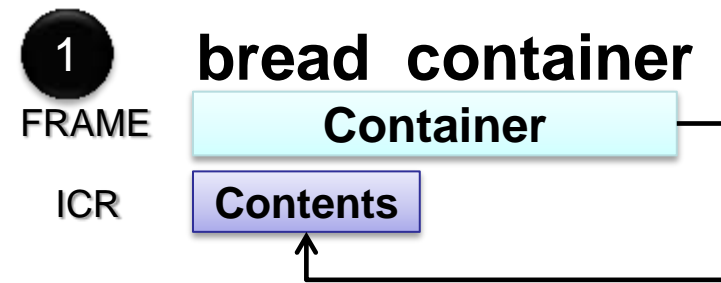


Experiments on Paraphrases

- Same setup as baseline relying on Semafor annotations on paraphrases
- Variables:
 - All paraphrases (high recall) vs. best rated paraphrase
 - Most frequent annotation vs all annotations
- Best results for „all“ paraphrases:
 - There are annotations for 60% of the Gold Standard
 - There is a frame-match for 148 MWEs
 - There is a role match for 42 MWEs (0.28 accuracy)
- Next steps:
 - Evaluate full sentences as paraphrases
 - Focus on precision by filtering paraphrases and/or annotations for constituents
 - Increase dataset and train classifier (next slide)

Acquisition of MWE Semantics using Paraphrases - revisited

- How to automatically discover ICRs?



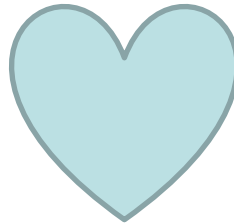
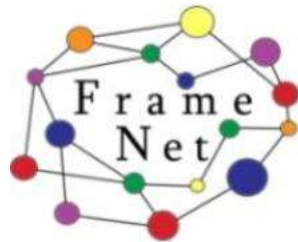
- ➔ More complex dependencies for certain paraphrases

More Paraphrases

- Semeval Shared 2013 Task for Generating Paraphrases of Noun Compounds is ongoing...
- Alternative: Generate Paraphrases from Parallel Text
 - Monolingual Translation Tables
 - Multilingual Translation Tables: EN→DE & DE→ EN leads to EN paraphrases („Pivoting“)
- Resource based on Pivoting: PPDB („The Paraphrase Database“), Ganitkevitch et al. (2013)
 - Overlap with Gold Standard MWEs: 75 Paraphrases in PPDB
 - Paraphrases mainly nominal phrases, of similar length
 - Example: travelling companion → travel companion, best partner
 - Example: waste material → waste materials, solid waste, waste disposal, hazardous wastes, materials that are

Summary/Outlook

- Motivation of MWE acquisition and ICR detection for FrameNet
- Basic setup for automatic ICR detection
 - Preliminary results
 - Difficult task requires a more complex setup



MWE

Multiword Expression

- Future work includes
 - Evaluating more complex classification setup and additional sources of paraphrases
 - Exploring the connection to relation detection for noun compounds (feature for ICR detection?)
 - Exploring methods of generalizing patterns

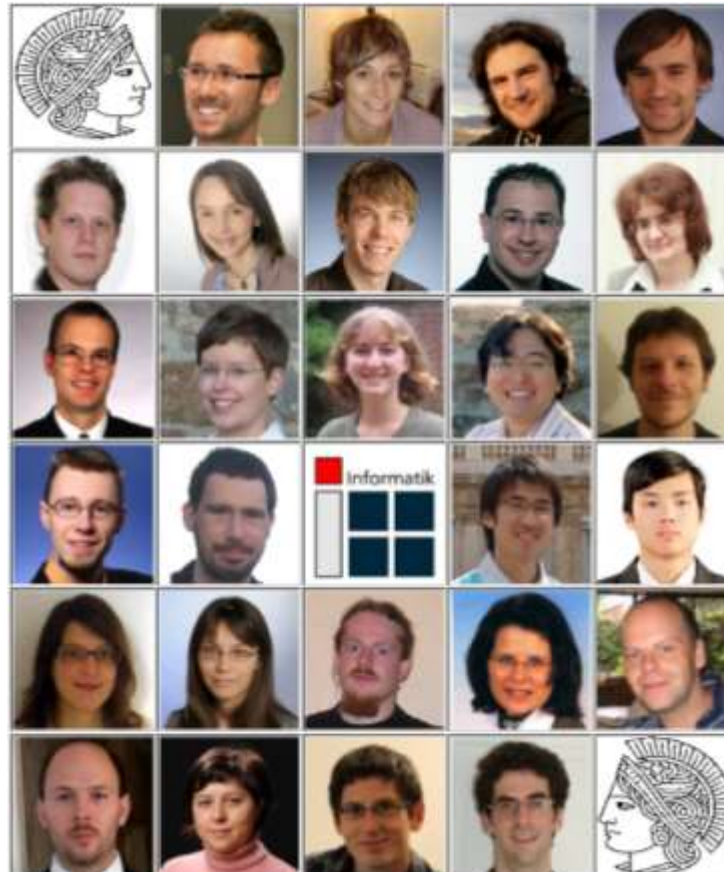
Thank you for your attention!



Ubiquitous Knowledge Processing Lab

 **LOEWE** – Landes-Offensive zur
Entwicklung **Wissenschaftlich-
ökonomischer Exzellenz**

 **DIPF**
Educational Research
and Educational Information



References

- Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A.. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pages 1934–1940, Las Palmas, Canary Islands, Spain.
- Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., & Scheffczyk, J. (2010). FrameNet II: Extended Theory and Practice. Berkeley, California: International Computer Science Institute, 2010.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, pages 1-15, London, UK, UK: Springer-Verlag.
- Erk, K., & Pado, S. (2006). Shalmaneser - A Toolchain For Shallow Semantic Parsing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, volume 6, pages 527-532, Genua, Italy.
- Pennacchiotti, M., De Cao, D., Basili, R., Croce, D., & Roth, M. (2008). Automatic induction of FrameNet lexical units. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 457-465, Honolulu, Hawaii.
- Tonelli, S., Giuliano, C., & Timoshenko, K. (2013). Wikipedia-based WSD for multilingual frame annotation. *Artificial Intelligence*, (194), 203-221.
- Laparra, E. & Rigau, G. (2010). Extended WordFrameNet. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pages 1214/1219, Valletta, Malta.

References

- Nakov, P. (2008). Paraphrasing Verbs for Noun Compound Interpretation. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), pages 46-49, Marrakech, Morocco.
- Butnariu, C., Kim, S. N., Nakov, P., Séaghdha, D. Ó., Szpakowicz, S., & Veale, T. (2010). SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pages 39-44, Stroudsburg, PA, USA.
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C.M., & Wirth, C. (2012). UBY – A Large-scale Unified Lexical Semantic Resource Based on LMF. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580--590, Avignon, France.
- Hartmann, S. & Gurevych, I. (2013): FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In: *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages to appear.
- Das, D., Schneider, N., Chen, D., & Smith, N.A. (2010). Probabilistic Frame-Semantic Parsing. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 948-956, Los Angeles, CA, USA.