UNIVERSITY OF CAMBRIDGE

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

University of Stuttgart
Germany

Predicting the Direction of Derivation in English Conversion

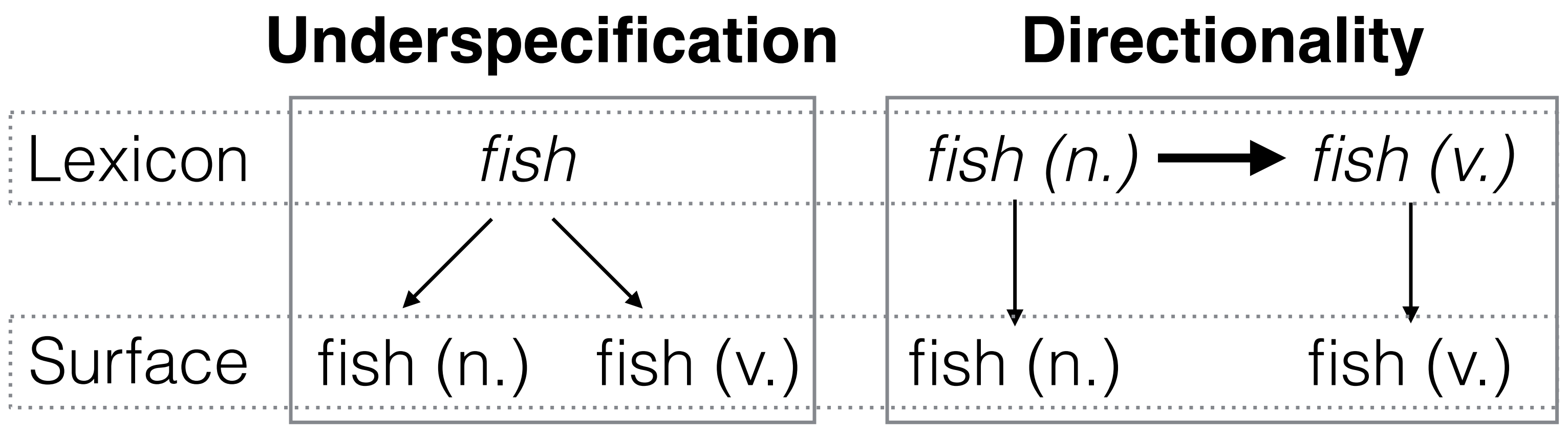
Max Kisselew* Laura Rimell† Alexis Palmer‡ Sebastian Padó*

* IMS, Stuttgart University, Germany † Computer Laboratory, University of Cambridge, UK

‡ Leibniz ScienceCampus, ICL, Heidelberg University, Germany

1. Morphology and Direction of Conversion

- **Conversion** changes grammatical category of a word without overt morphological marking, e. g.: *tunnel (n.)* → *tunnel (v.)*, *walk (v.)* → *walk (n.)*
- Various theoretical accounts of conversion: Uncategorized roots (underspecification) vs. directed derivation



Research Question

In a corpus-based study, which factors are able to account for diachronic precedence in cases of English V-to-N and N-to-V conversion?

2. Hypotheses

1. Derived forms are **less frequent** than their bases (Harwood and Wright, 1956; Hay, 2001)
2. Derived forms are **more semantically specific** than their bases (Koontz-Garboden, 2007; Plag, 2003), as approximated by information theoretic measures

3. Data

- **Gold standard**: Historical precedence data from CELEX (Baayen et al., 1995) for English
  - 1,044 monomorphemic English N-to-V lemma pairs
  - 948 monomorphemic English V-to-N lemma pairs
- **Corpus**: Concatenation of the lemmatized and part-of-speech (PoS) tagged BNC and ukWaC corpora containing 2.36 billion tokens
- **Semantic vector space**: Separate vectors *c.noun* and *c.verb* for each conversion case *c*
  - BOW count vectors, 10000 dimensions, context window ±5
  - Downsampling: For each verb-noun conversion pair, both vectors are constructed from the same number of occurrences

4. Specificity Measures

- Two measures for semantic specificity of a word:
  - Entropy:
    $H(\boldsymbol{v}) = -\sum_{i \in \boldsymbol{v}} \boldsymbol{v}_i \cdot log(\boldsymbol{v}_i)$
    (high semantic specificity ∼ low entropy)
  - Kullback-Leibler (KL) divergence:
    $D(\boldsymbol{v}||\boldsymbol{n}) = \sum_{i} \boldsymbol{v}_i \cdot log(\frac{\boldsymbol{v}_i}{\boldsymbol{n}_i})$
    (high semantic specificity ∼ high KL divergence from neutral vector)
    - KL divergence between term vector and “neutral” context vector *n* as a measure of the vector’s semantic specificity
    - Here: “neutral” vector *n* computed as centroid vector for all words in the corpus

5. Experiments

- Testing hypothesis 1 (Frequency):
  If f(N) > f(V) then N-to-V (else V-to-N)
- Testing hypothesis 2 (Semantic specificity):
  If H(N) > H(V) then N-to-V (else V-to-N)
  If D(V||n) > D(N||n) then N-to-V (else V-to-N)
  (where *n* is the neutral vector)
- Combined model: combination of individual indicators (standardized differences in log frequency, entropy, and KL divergence within each pair) as features in a logistic regression model

6. Results

Predictor	N-to-V	V-to-N	all
Most Freqent Class	100%	0%	52.4%
Entropy *H*	50.1%	75.5%	62.2%
KL divergence	53.8%	**76.7%**	64.6%
Frequency	**84.7%**	58.7%	72.3%
Freq + *H* + KL	77.4%	76.0%	**76.8%**

Accuracies for predicting the direction of derivation

- Large difference in results between N-to-V and V-to-N
- Frequency best predictor for N-to-V cases
  - Large variety in meaning shifts
  - Verb describes an ‘action having to do with the noun’. E. g.: *celluloid the door open*, meaning ‘use a credit card to spring the lock open’ (Clark and Clark, 1979)
  - Irregular semantics of conversion
- Specificity predictors better for V-to-N cases
  - Noun is likely to refer to the event described by the verb or its result (Grimshaw, 1990)
  - More regular semantics of conversion
- Simple combination does well for both cases

7. Discussion and Conclusion

- Striking complementarity in the ability of frequency and semantic specificity to account for the direction of conversion in N-to-V and V-to-N cases
- N-to-V conversion consistent with underspecification approach
- V-to-N conversion consistent with directionality approach

Acknowledgements

MK and SP acknowledge partial funding by Deutsche Forschungsgemeinschaft (SFB 732, Project B9). LR acknowledges EPSRC grant EP/I037512/1 and ERC Starting Grant DisCoTex (306920). AP acknowledges Leibniz Association grant SAS-2015-IDS-LWC and the Ministry of Science, Research, and Art of Baden-Württemberg.

References

Baayen, H. R., R. Piepenbrock, and L. Gulikers (1995). *The CELEX lexical database. Release 2. LDC96L14*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Clark, E. V. and H. H. Clark (1979). When nouns surface as verbs. *Language 55*, 767–811.

Grimshaw, J. (1990). *Argument Structure*. Cambridge: MIT Press.

Harwood, F. W. and A. M. Wright (1956). Statistical study of English word formation. *Language 32*(2), 260–273.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics 39*, 1041–70.

Koontz-Garboden, A. (2007). *States, changes of state, and the Monotonicity Hypothesis*. Ph. D. thesis, Stanford University.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.