Building a Semantic Representation for German

Midterm Presentation

University of Bonn

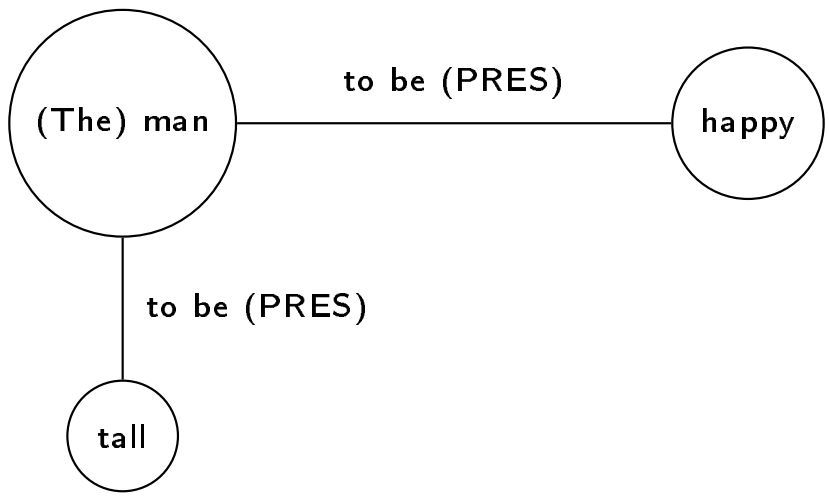
12.06.2018

Motivation

- preparation for a bachelor thesis related to machine translation
- build a graph-based meaning representation (MR)
  - as 'interlingua' for machine translation
  - to allow more sophisticated search in texts
- build a basic framework
  - to parse german sentences into the MR
  - to generate german sentences from MR graphs
- in the bachelor thesis:
  - extend the framework to another language
  - build a basic machine translation system which uses the intermediate MR

Example

(1) The man who is tall is happy.



Semantic Representations

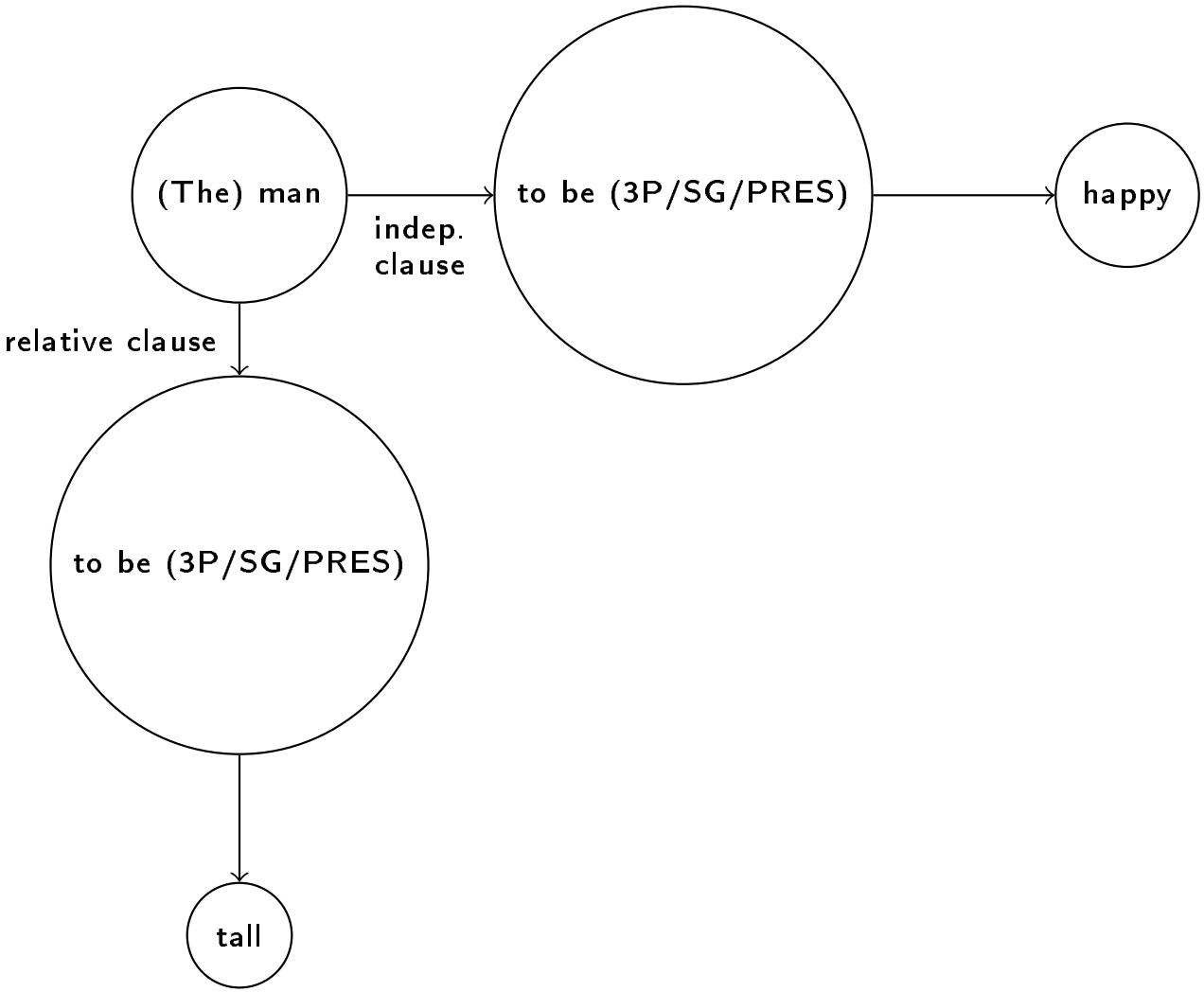
- one of the first approaches was the Montague grammar
  - developed in the 60s and early 70s
  - based on formal logic
- a lot of other representations have been developed:
  - conceptual meaning representation
  - thematic role representations
  - first order logic
  - discourse representation theory
  - semantic networks

My Semantic Representation

- graph structure (labeled nodes and edges)
- it should indicate verbs and their arguments
- it should indicate to what a prepositonal phrase refers
- it should indicate which parts are expressed in independent clauses/dependent clauses
  - e.g. to which noun does a relative clause refer?
- it should be as expressive as the natural language sentence
  - all information expressed has to be represented
  - should be possible to generate at least 'an equivalent' (or better: the same) sentence from the graph

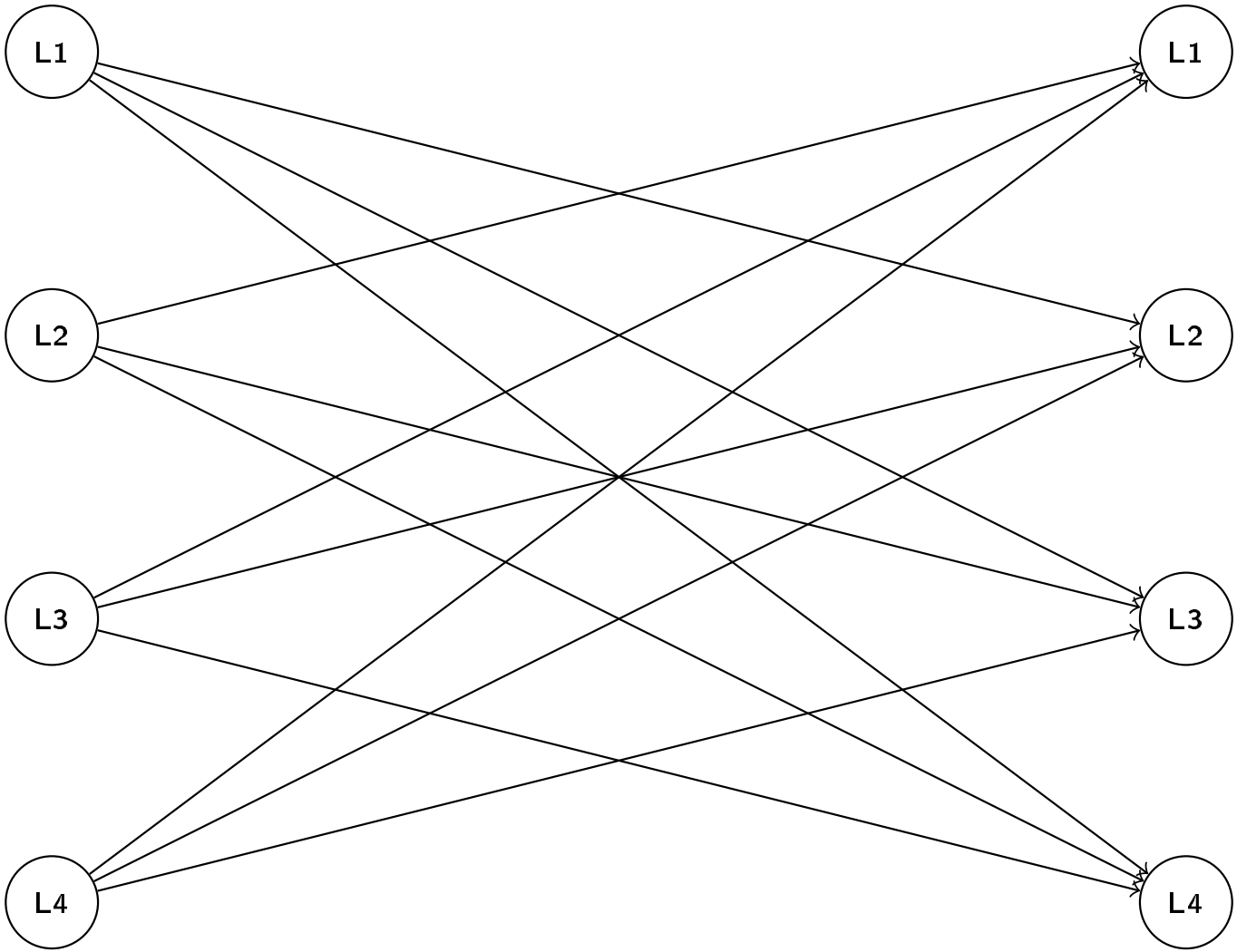
Better Example

(1) The man who is tall is happy.



Standard Approach

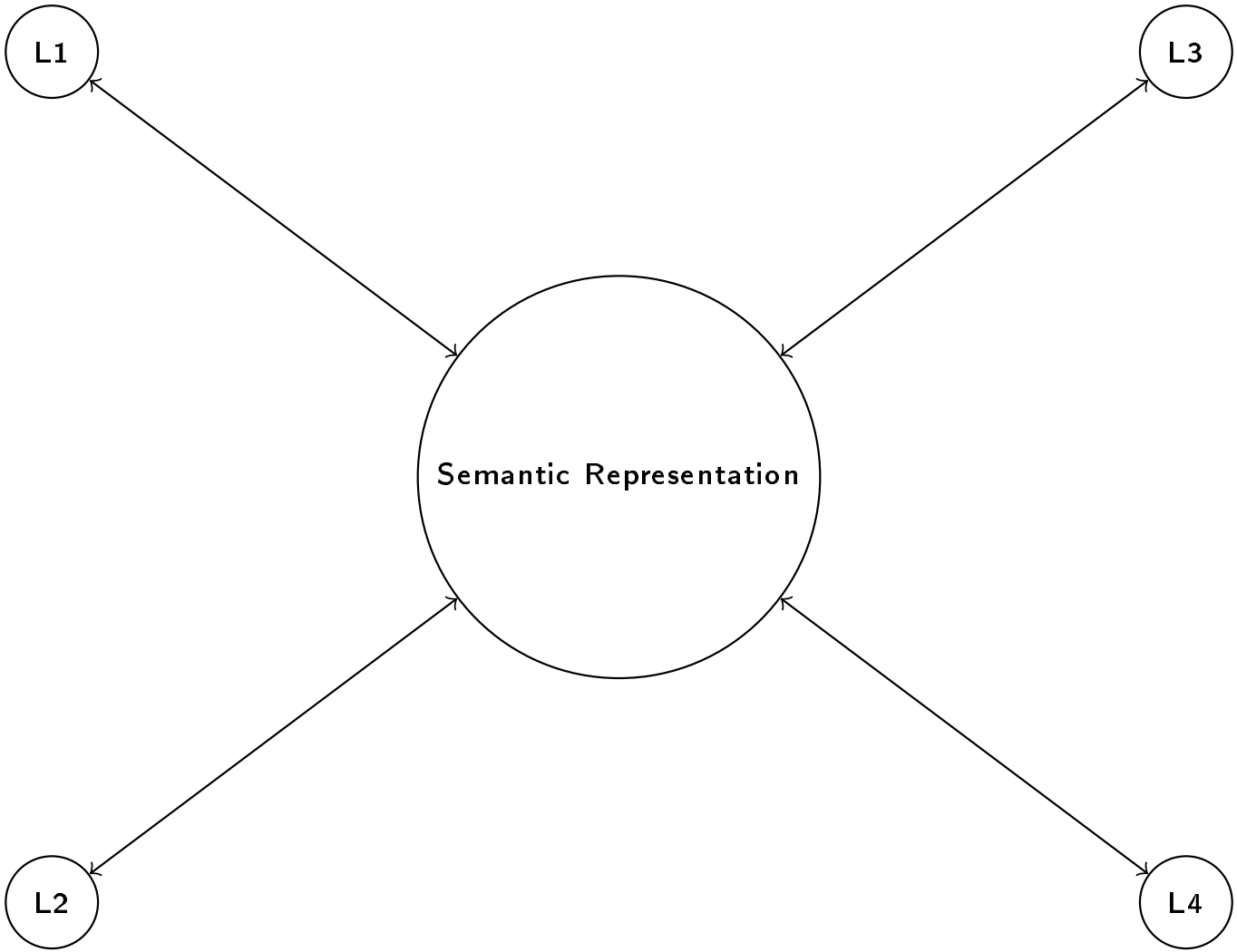
Source/Target Language Pairs



- for n languages one has $n(n-1)$ source/target language pairs

Interlingual Approach

Source/Target Language Pairs



- for n languages one has $2n$ source/target language pairs

Interlingual Approach

Advantages and Problems

- Advantages:
  - obviously less source/target language pairs
    - therefore allows more sophisticated, even (partly) hand-coded rule-based models
  - no need for parallel corpora to train
    - therefore translation for fancy pairs would be possible
- Problems:
  - how 'universal' can a representation be without losing expressivity?
  - a model for a interlingua is likely to be strongly influenced by the authors native language

Ambition

- will this project result in a state-of-the-art representation for interlingual machine translation that will be widely used?
  - probably not
  - rather it will (hopefully) be useful as an imperfect representation for German, English, Romance languages and maybe some other Indo-European languages
  - furthermore, it is an interesting project to explore a wide variety of machine translation problems, such as
    - lemmatization/stemming
    - POS tagging
    - syntactical analysis
    - semantic analysis of sentences

Tentative Roadmap

- write a lemmatizer/stemmer for german
  - the NLTK 'Snowball Stemmer' is horrible
  - spaCy is better, but still insufficient
- train a POS tagger/syntax parser using the lemmatization tool on the TIGER Corpus
- build semantic representation graphs from the syntax trees
  - will probably work rule based
- build sentences from semantic representation graphs
  - will probably work rule based as well
  - I will focus on 'simple sentences' first and try to get as far as possible

Questions?

References

- Schubert, Lenhart: Semantic Representation. In: Proceedings of the Twenty-Ninth (AAAI) Conference on Artificial Intelligence, p. 4132-4139 (2015)
- Jurafsky and Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River, New Jersey: Prentice Hall, 2000. Print (Chapter on Machine Translation: p.799-831)