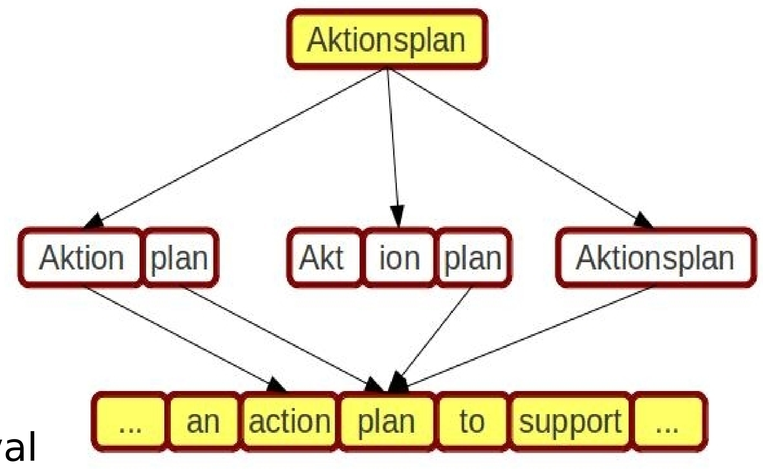
# Compound Splitting for Statistical Machine Translation

Bilha Marindi, Hobli Taffame, Zarina Soltobaeva
Department of Computational Linguistics
Heidelberg University, Germany

**Abstract** The goal of this work is to split compounds in meaningful words, which leads to a better machine translation. Compound words appear in many languages as: This work is based on German-English Europarl-Corpus.



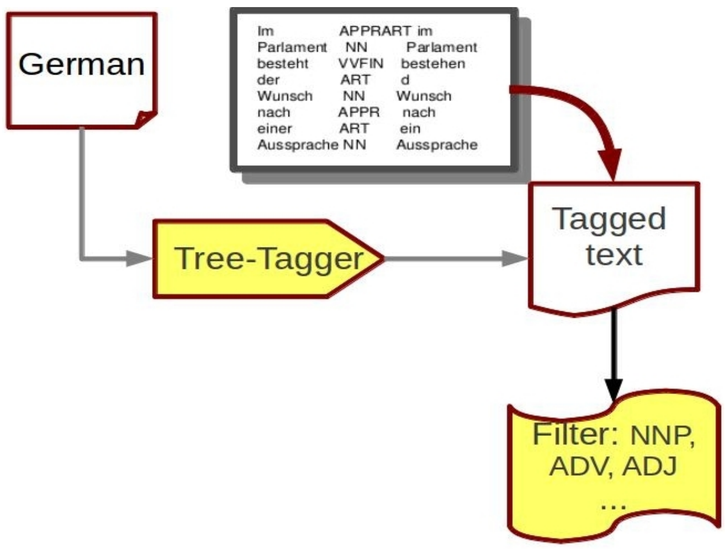

## Methods

1. Frequency Based Metrics
2. Limitation on Part-of-Speech
3. Parallel Corpus

## Usage

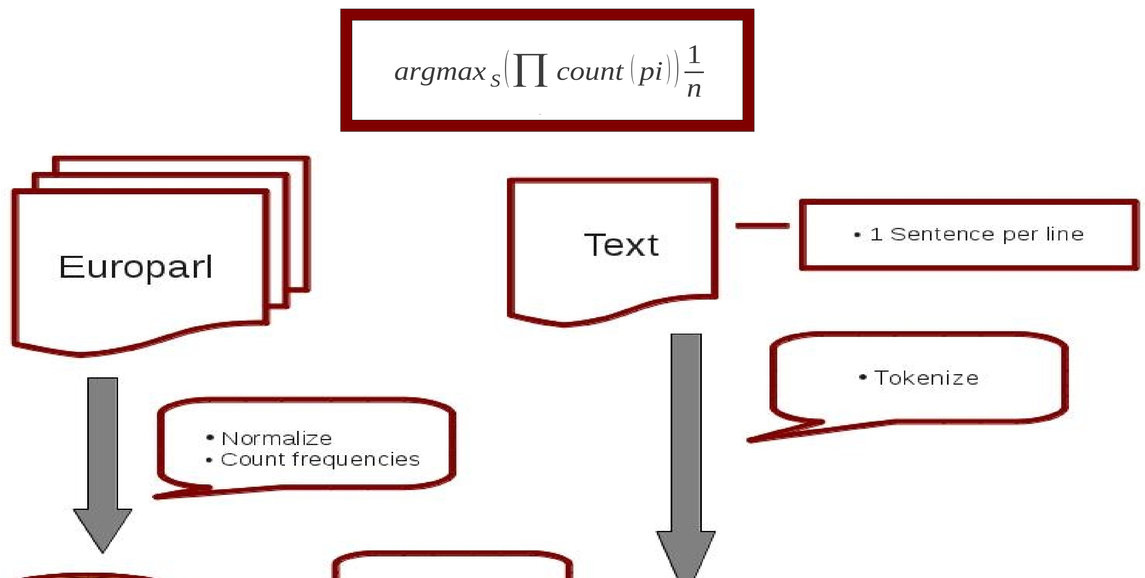
Machine Translation
Speech recognition
Text classification
Infromation extraction or retrieval

## Limitation on Part-of-Speech

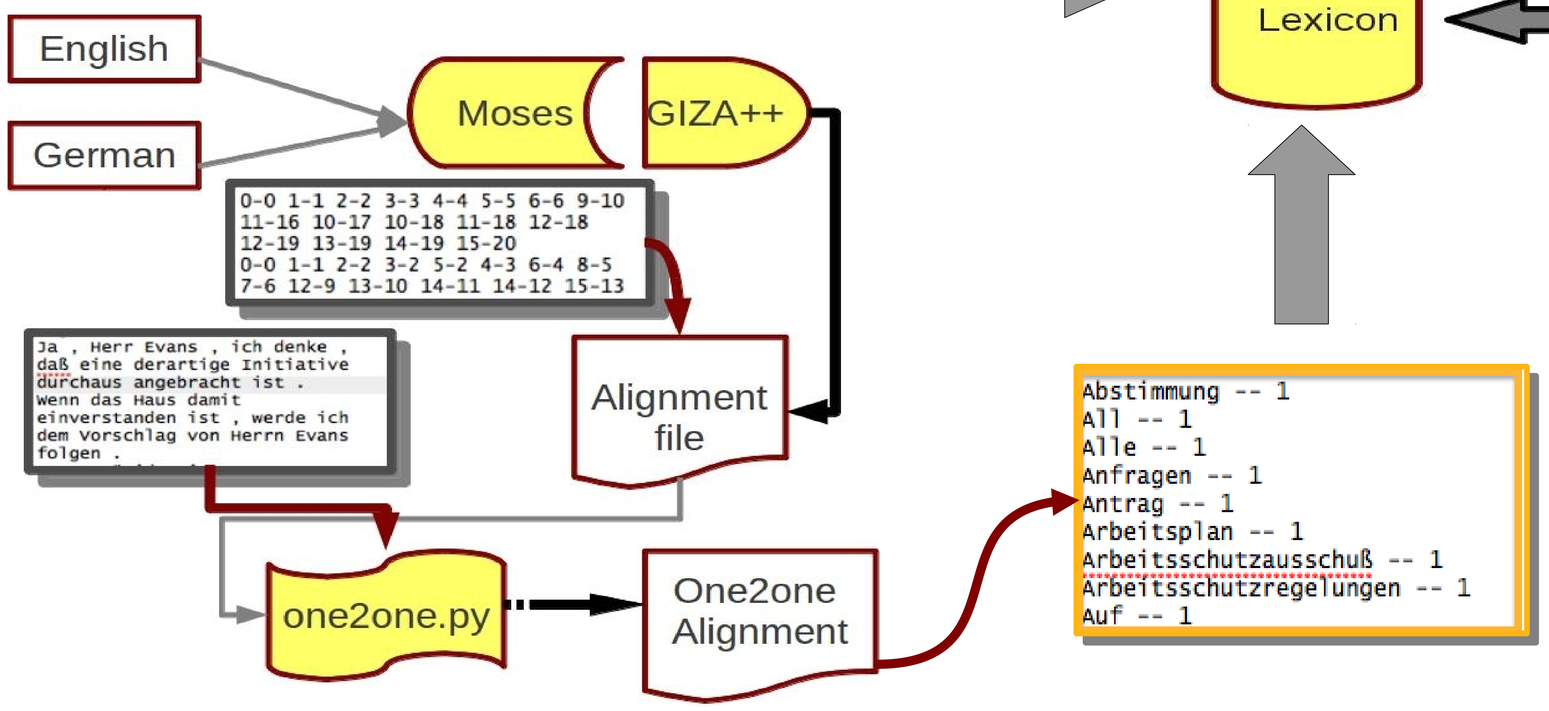



## Frequency Based Metrics

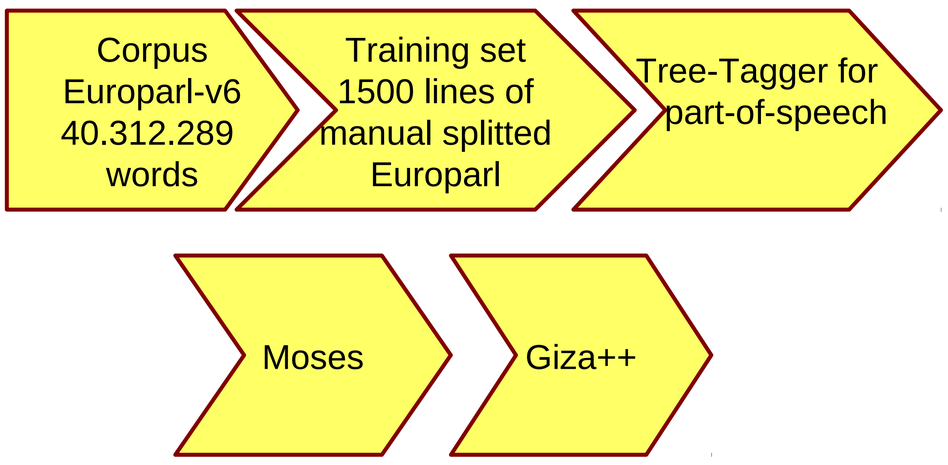
$$argmax_S\left(\prod count\,(pi)\right)\frac{1}{n}$$




## Parallel Corpus




## Resources




## Example

Wiederaufnahme
wie(142660) -der(1489558) -aufnahme(3896) → 93898.68
wieder(20378) -aufnahme(3896) → 8910.25

## Evaluation

| Method | Metrics | |
|---|---|---|
| | Precision | Recall |
| Frequency Based Metrics | 49.3% | 50.7% |
| Limitation on Part-of-Speech | 58.5% | 41.5% |
| Parallel Corpus | 49.6% | 40.4% |
| Parallel Corpus with POS | 49.2% | 40.8% |

## Reference

P. Koehn, K. Knight. *Empirical Methods for Compound Splitting*