# Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation

Nadir Durrani

Philipp Koehn

# Outline

- Spoken Language Translation
  - Confusion Networks
  - Punctuation
- Machine Translation
  - Operation Sequence Models
  - Word classes (Brown clusters, POS and Morph tags)

# SLT

- Match ASR output to MT input
  - Punctuation before translation (Matusov et al. 2006)
- Leverage uncertainty in ASR output
  - Confusion Network (Bertoldi et al. 2007)

Maximise the benefit from using Edinburgh ASR systems

# SLT

- English ASR System
  - Combines tandem & hybrid DNN acoustic model
  - Speaker adaption on test set
  - Recurrent NN LM

"The UEDIN English ASR System for the IWSLT 2013 Evaluation", Peter Bell, Fergus McInnes, Siva Reddy Gangireddy, Mark Sinclair, Alexandra Birch, Steve Renals.

# SLT

- German ASR System
  - KALDI toolkit
  - Hybrid DNN with 6 hidden layers, and 2048 nodes
  - 4-gram LM

"Description of the UEDIN System for German ASR", Joris Driesen, Peter Bell, and Steve Renals.

# SLT

- Baseline Systems
  - MOSES Phrase-Based Model
  - TED + Large out-of-domain corpora (Europarl, News Commentary, Multi UN, Gigaword, and Common Crawl)
  - Domain Filtering:

    bilingual cross-entropy (Axelrod et al. 2011)
  - German source:

    compound splitting (Koehn and Knight 2003) and syntactic pre-ordering (Collins et al 2005)

# SLT

| Model Type | tst2010 |
|---|---|
| All data | 30.8 |
| Domain filtering: bilingual cross entropy | 31.6 $\boxed{(+0.8)}$ |
| Filtering + strip source punctuation | 28.4 $\boxed{(-3.2)}$ |

10% Out of Domain

English - French:
Baseline system trained and test on transcripts

# SLT

| Model Type | tst2010 | |
|---|---|---|
| All data | 21.4 | |
| Domain filtering: bilingual cross entropy | 27.8 (+6.4) | 20% Out of Domain |
| Filtering + no syntactic preordering | 24.3 (-3.5) | |
| Filtering + no syntactic preordering + strip source punctuation | 23.6 (-4.2) | |

German - English:

Baseline system trained and test on transcripts

# SLT

- SLT experiments with ASR input
- Experimental setup:
  - Lattices reduced, and remove nulls
  - Lattices also truecased and tokenised:

    Europe’s $\rightarrow$ Europe ’s
  - Prune phrase table: 402 000 translations of “a”
  - Punctuation MT model: strip punctuation on source side, monotone decoding

# SLT

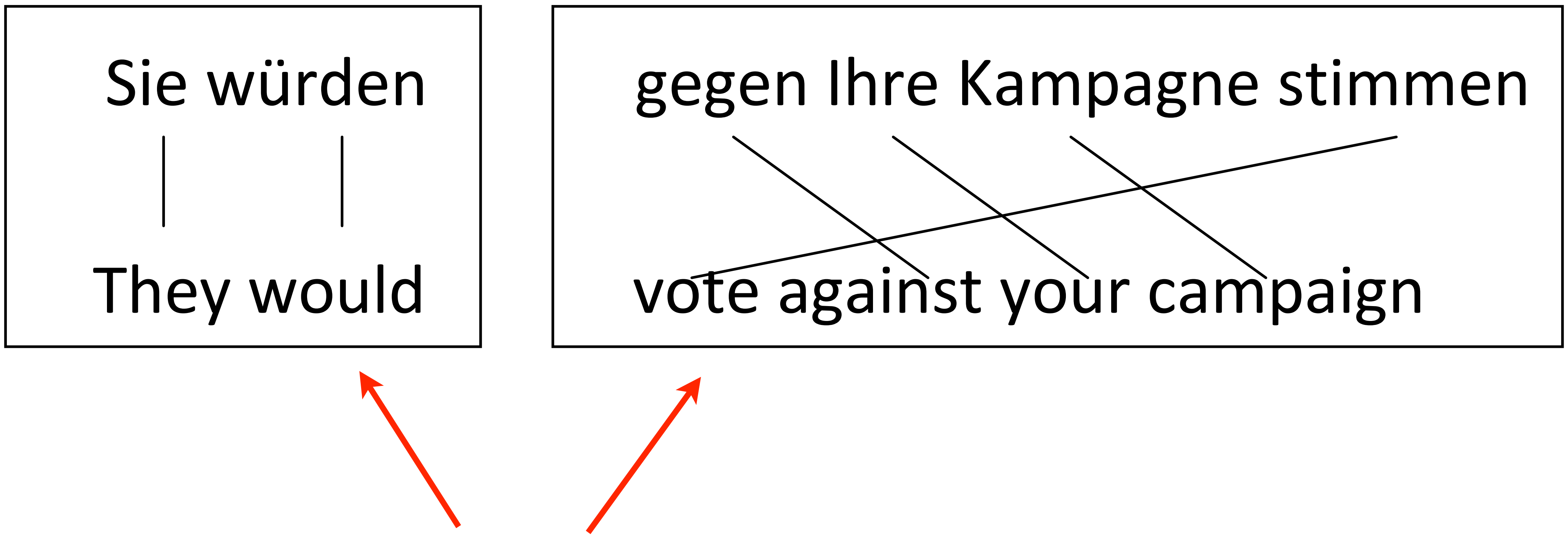
| Model Type | en-fr (tst2010) | de-en (dev2012!) |
|---|---|---|
| Absolute 1-best | 22.9 | 17.0 |
| Absolute 1-best Punctuated | 24.1 (+1.2) | 16.1 (-0.9) |
| Lattice 1-best | 17.9 (-5.0) | - |
| Confusion Network | 19.5 (-3.4) | 11.1 (-5.9) |

WER: 17.0 18.6

# MT

- All languages + English as source or target!
- Sequence models (language models) and operation sequence models (OSMs) over generalised representations:
  - Brown clusters
  - POS tags
  - Morphological tags

# Phrase-based Models




Strong independence assumption

# Phrase-based Models

gegen Ihre Kampagne stimmen

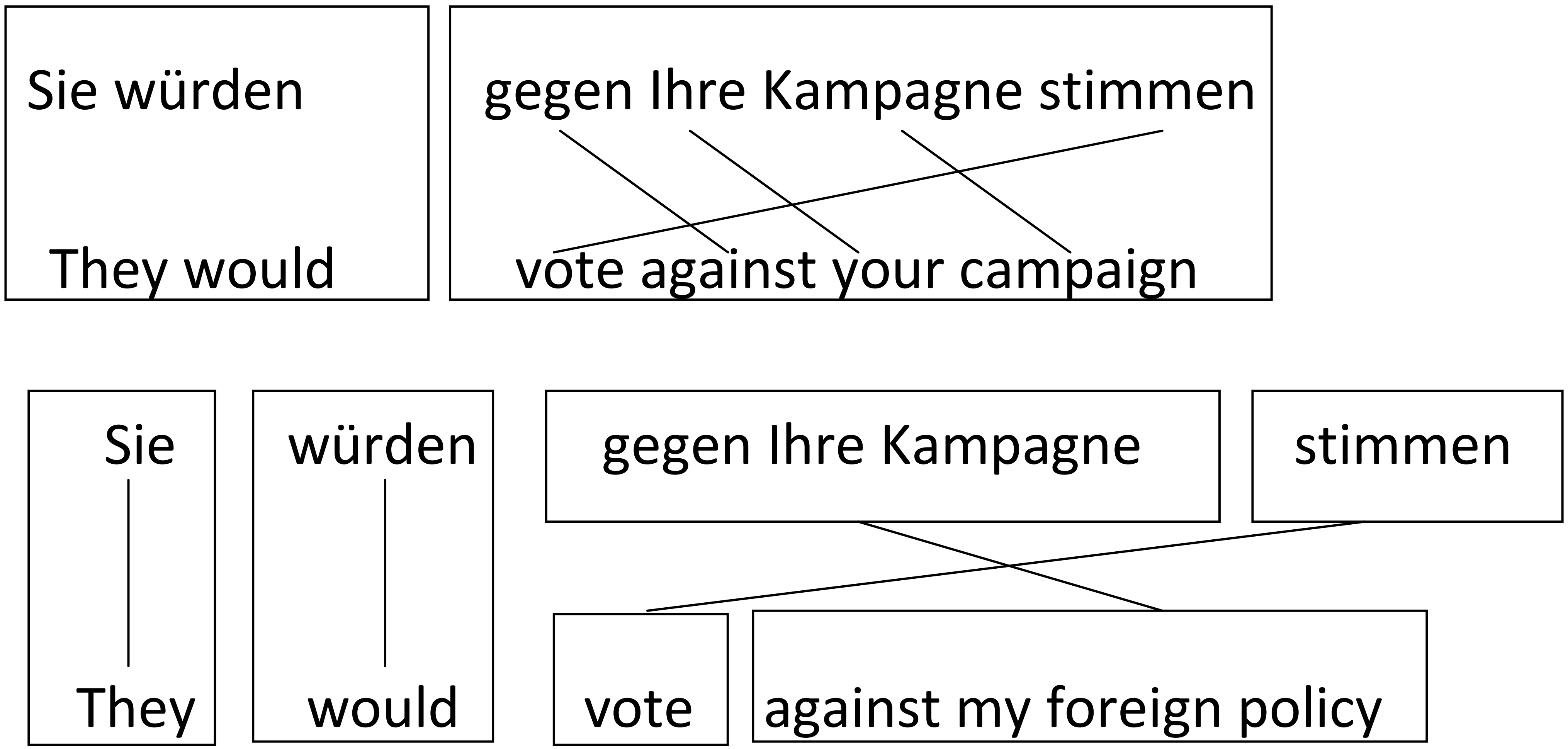
vote against your campaign

gegen sie stimmen

vote against you

doesn't generalise

# Phrase-based Models




spurious segmentations

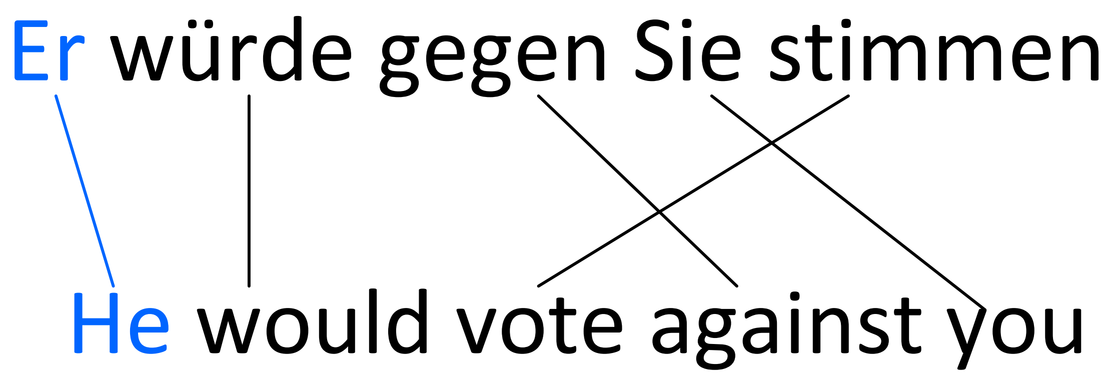
# Operation Sequence Model

- Instance of N-gram based SMT framework (Marino et a. 2006)
- Translation as a sequence of operations (Durrani et al. 2011)
  - Strongly integrates translation and reordering operations

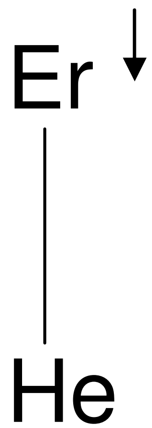
$$p_{osm}(F, E, A) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, \ldots, o_{j-1})$$

Markov chains

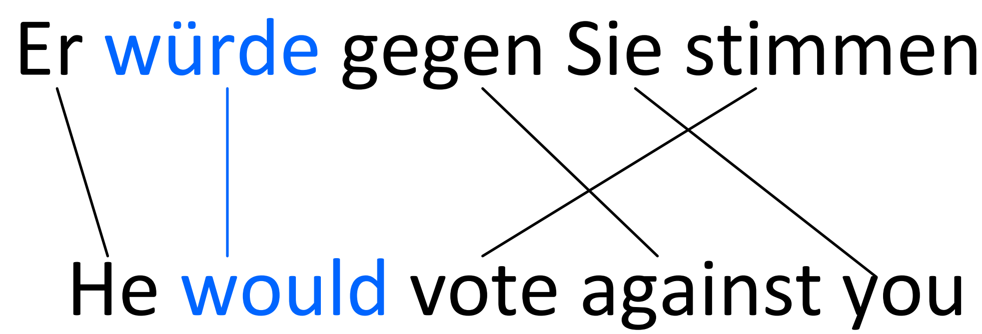
# Operation Sequence Model



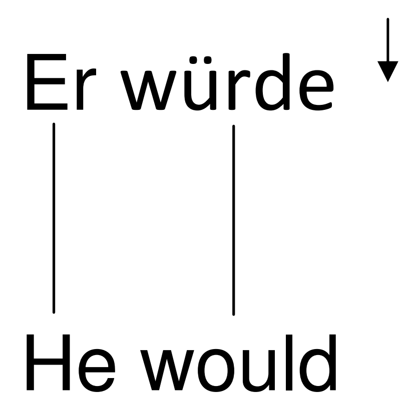

- Operations
  - $o_1$: Generate (Er – He)



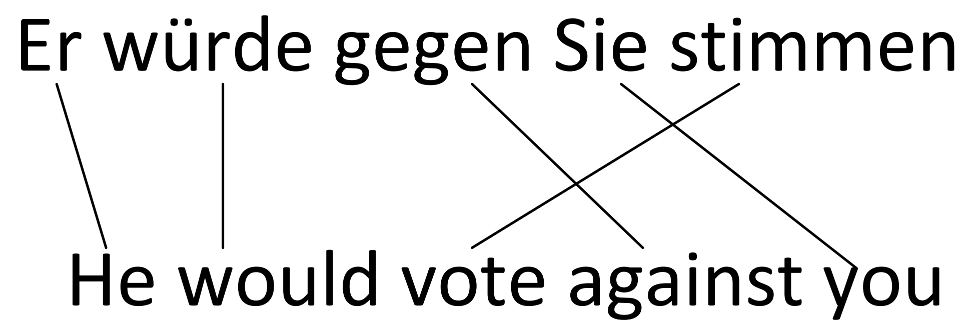
# Operation Sequence Model



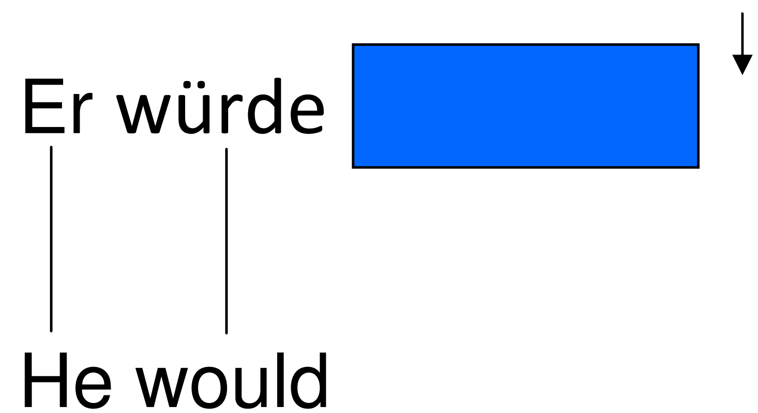

- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)



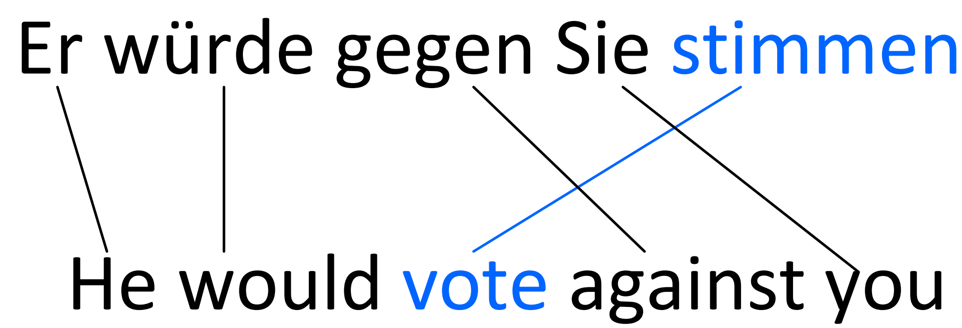
# Operation Sequence Model



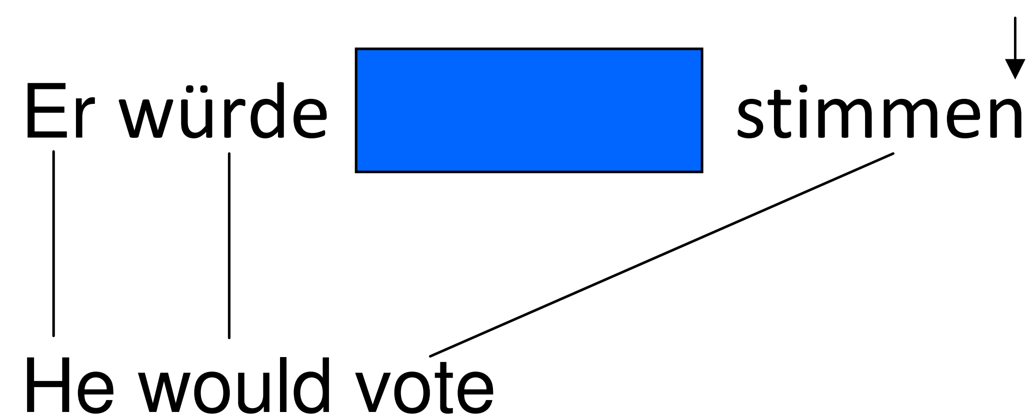

- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap



# Operation Sequence Model




- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap
  - $o_4$ Generate (stimmen, vote)



# Operation Sequence Model

Er würde gegen Sie stimmen

He would vote against you

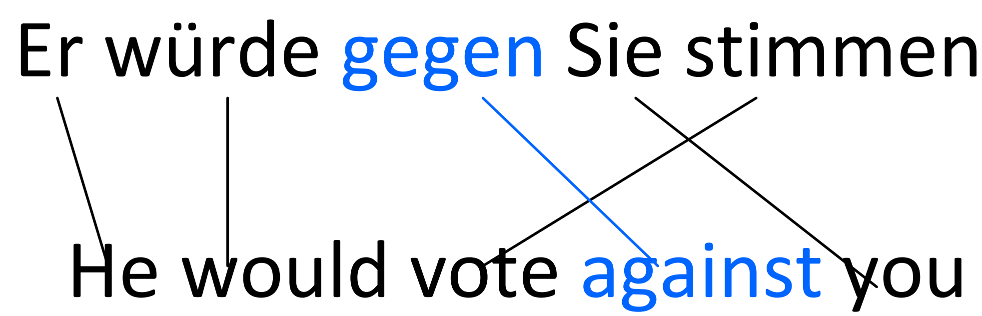
- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap
  - $o_4$ Generate (stimmen, vote)
  - $o_5$ Jump Back (1)

↓

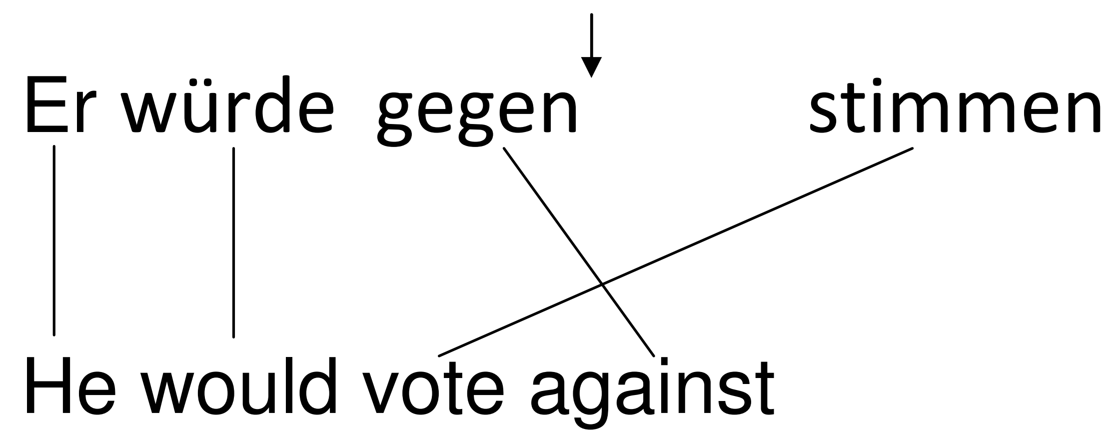
Er würde stimmen

He would vote

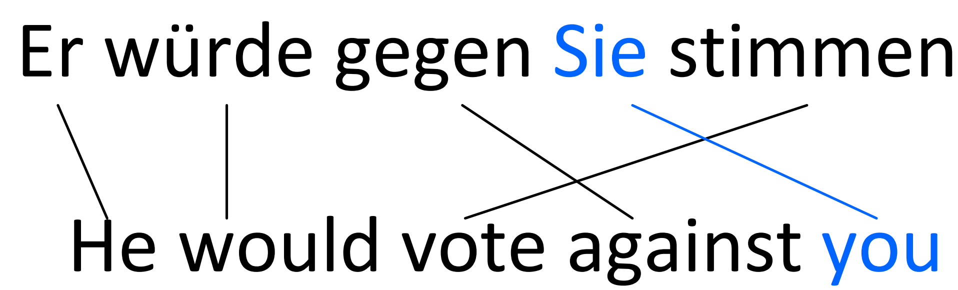
# Operation Sequence Model



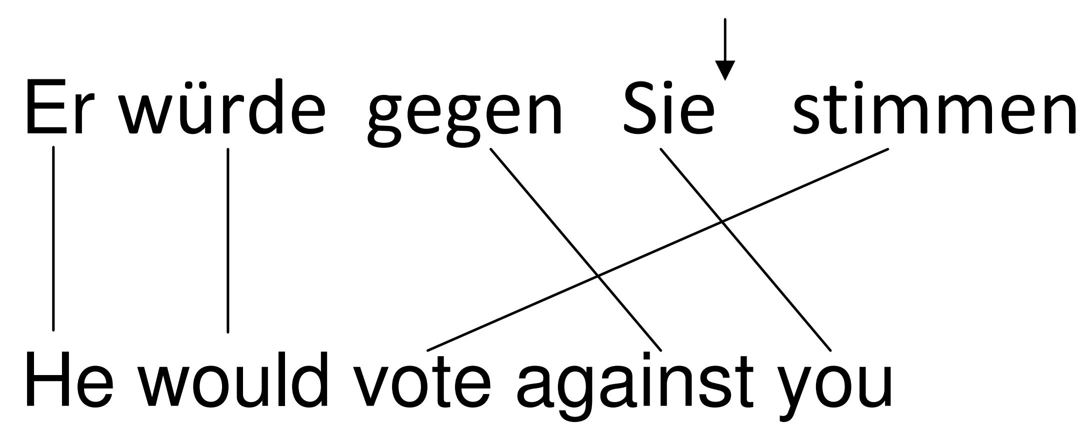

- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap
  - $o_4$ Generate (stimmen, vote)
  - $o_5$ Jump Back (1)
  - $o_6$ Generate (gegen, against)



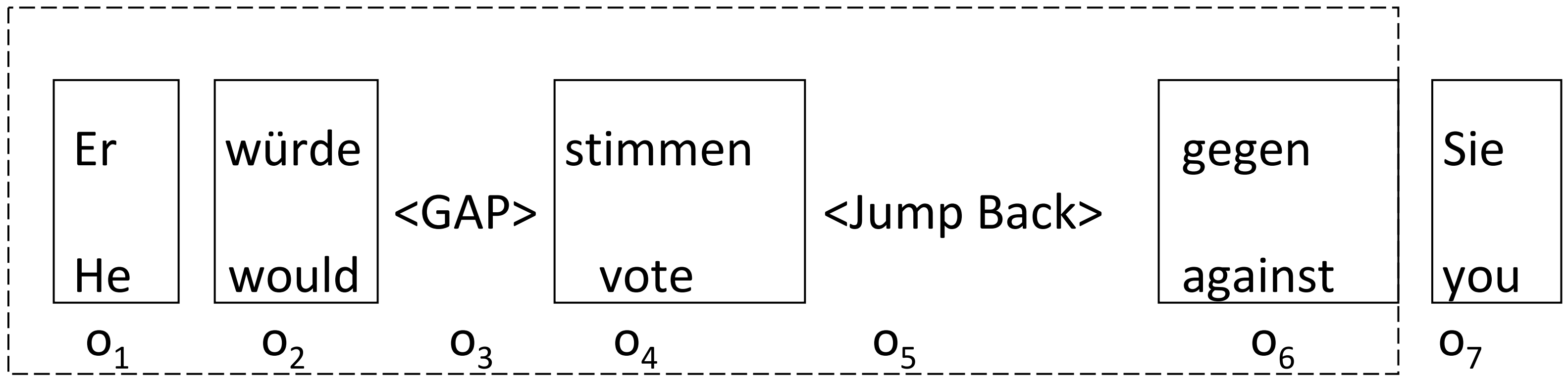
# Operation Sequence Model




- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap
  - $o_4$ Generate (stimmen, vote)
  - $o_5$ Jump Back (1)
  - $o_6$ Generate (gegen, against)
  - $o_7$ Generate (Sie, you)



# Operation Sequence Model




Context Window: 9-gram Model

- Markov model over operation sequence
  - Contextual information across phrase boundaries: does not make phrasal independence assumptions
  - Does not have spurious segmentation ambiguity
  - Memorises reordering patterns: consistently handles local and non local reorderings

# OSM for word classes

- OSM has shown to help overcome independence assumptions of phrase-based MT
- BUT: sparse so useful over small context
- Apply to generalised representations ← NEW
  - Sparse counts for operations over words
  - Extend context by using Brown clusters etc.
  - More robust statistics

# MT Baselines

- MOSES phrase-based model
  - OSM model features over words
  - sparse features: domain indicator, lexical, phrase length, and count bin
  - factored models for German–English and English–German
  - hierarchical lexicalized reordering (mslr)
  - MADA tokenizer for source-side Arabic
  - Stanford Chinese segmenter

# MT Baselines

| Language | Into English | From English |
| --- | --- | --- |
| Arabic | 24.8 | 7.6 |
| Chinese | 11.8 | 9.8 |
| Dutch | 32.8 | 26.5 |
| Farsi | 14.5 | 8.0 |
| French | 33.3 | 33.2 |
| German | 30.5 | 22.9 |
| Italian | 29.7 | 23.7 |
| Polish | 17.7 | 9.7 |
| Portuguese | 36.0 | 30.8 |
| Romanian | 31.7 | 21.1 |
| Russian | 19.1 | 13.1 |
| Slovenian | 24.7 | 18.0 |
| Spanish | 39.5 | 33.9 |
| Turkish | 13.5 | 7.2 |

tst2010

# Brown Clusters

- Brown clusters (Mediani et al. 2012, Wuebker 2013, Ammar et al. 2013)
  - Motivation: lack of tools for many languages
  - word classes that are optimised to reduce n-gram perplexities

    group adjectives with same inflection - syntax

    group all colour adjective - semantics
  - Use GIZA++ mkcls on the target side of parallel corpus
  - Train an n-gram sequence model over these identifiers as an additional scoring function

# Brown Clusters

- Brown clusters better than POS + Morph tags?
    - Brown clusters are more evenly distributed than POS tags where the distribution is biased toward the noun classes
    - Brown clusters are optimised for language modelling
    - Using POS and morph tags can increase sparsity because you can assign a word different tags

# LM over Brown clusters

| Language | $B_0$ | 50 | 200 | 600 | 1000 |
|---|---|---|---|---|---|
| Dutch | 26.5 | **26.7** | 26.2 | 26.3 | 26.5 |
| | | +0.2 | –0.4 | –0.2 | ±0.0 |
| French | 33.2 | 33.3 | **33.4** | 33.1 | 33.1 |
| | | +0.1 | +0.2 | –0.1 | –0.1 |
| Polish | 9.7 | 9.9 | 10.1 | 10.1 | **10.4** |
| | | +0.2 | +0.4 | +0.4 | +0.7 |
| Portuguese | 30.8 | 31.6 | 32.2 | 32.4 | **32.4** |
| | | +0.8 | +1.4 | +1.6 | +1.6 |
| Russian | 13.1 | 13.3 | 13.5 | 13.5 | **14.0** |
| | | +0.2 | +0.4 | +0.4 | +0.9 |
| Slovenian | 18.0 | **18.7** | 18.6 | 17.7 | 18.0 |
| | | +0.7 | +0.6 | –0.3 | ±0.0 |
| Spanish | 34.1 | 34.3 | **34.6** | 34.5 | 34.0 |
| | | +0.2 | +0.5 | +0.4 | –0.1 |
| Turkish | 7.2 | 7.4 | 7.5 | **7.5** | 7.5 |
| | | +0.2 | +0.3 | +0.3 | +0.3 |

# OSM over Brown clusters

| Language | $B_0$ | 50 | 200 | 600 | 1000 |
|---|---|---|---|---|---|
| Dutch | 26.5 | **26.9** +0.2 | 26.5 +0.3 | 26.6 +0.3 | 26.5 ±0.0 |
| French | 33.2 | **33.8** +0.5 | 33.7 +0.3 | 33.6 +0.5 | **33.8** +0.7 |
| Polish | 9.7 | 10.1 +0.2 | **10.2** +0.1 | **10.2** +0.1 | 10.1 –0.3 |
| Portuguese | 30.8 | 31.8 –.02 | 32.4 +0.2 | **32.3** –0.1 | 31.9 –0.5 |
| Russian | 13.1 | 13.6 +0.3 | 13.7 +0.2 | **13.8** +0.3 | 13.6 –0.4 |
| Slovenian | 18.0 | 18.6 –0.1 | **18.9** +0.3 | 18.2 +0.5 | 18.0 ±0.0 |
| Spanish | 34.1 | **34.7** +0.4 | 34.6 ±0.0 | 34.6 –0.1 | 34.6 +0.6 |
| Turkish | 7.2 | 7.3 –0.2 | 7.3 –0.2 | **7.5** ±0.0 | **7.5** ±0.0 |

# OSM over POS + Morph

| Model | English-German | German-English |
|---|---|---|
| **Baseline** | 22.9 | 30.5 |
| **+OSM$_{(pos,pos)}$** | 23.2 +0.3 | 31.0 +0.5 |
| **+OSM$_{(pos,morph)}$** | 23.9 +1.0 | 31.2 +0.7 |
| **+OSM$_{all}$** | 24.2 $\boxed{+1.3}$ | 31.1 $\boxed{+0.6}$ |
| | **English-French** | **English-Spanish** |
| **Baseline** | 33.1 | 33.9 |
| **+OSM$_{(pos,pos)}$** | 33.0 $\boxed{-0.1}$ | 34.4 $\boxed{+0.5}$ |
| | **English-Dutch** | |
| **Baseline** | 26.6 | |
| **+OSM$_{(pos,pos)}$** | 26.6 $\boxed{\pm 0.0}$ | |

# MT Official Submission

| Language | Into English test$_{11}$ | Into English test$_{12}$ | Into English test$_{13}$ | From English test$_{11}$ | From English test$_{12}$ | From English test$_{13}$ |
|---|---|---|---|---|---|---|
| Arabic | 25.6 | 27.7 | 26.3 | 11.9 | 12.4 | 11.5 |
| Chinese | 16.1 | 14.2 | 15.3 | 19.8 | 18.1 | 18.6 |
| Dutch | 36.0 | 33.0 | 32.7 | 30.3 | 26.7 | 25.5 |
| Farsi | 19.2 | 15.9 | 15.1 | 12.3 | 10.2 | 9.5 |
| French | – | – | – | 40.6 | 41.2 | 38.5 |
| German | – | – | 25.5 | 27.1 | 22.5 | 24.0 |
| Italian | 30.2 | 29.6 | 34.9 | 24.4 | 25.3 | 29.2 |
| Polish | 21.7 | 18.5 | 20.9 | 13.1 | 10.5 | 11.5 |
| Portuguese | 39.0 | 40.6 | 37.3 | 33.6 | 34.9 | 33.2 |
| Romanian | 36.1 | 31.8 | 29.8 | 23.2 | 19.2 | 17.6 |
| Russian | 22.1 | 20.7 | 22.7 | 15.9 | 13.5 | 16.1 |
| Slovenian | – | 21.2 | 24.1 | – | 12.4 | 13.7 |
| Spanish | 37.1 | 30.8 | 39.1 | 33.2 | 26.8 | 34.7 |
| Turkish | 15.0 | 15.0 | 14.9 | 7.4 | 7.4 | 6.8 |

# Summary

- SLT
  - Punctuation ASR input helps
  - WER is important for MT performance
  - Confusion Networks help
  - Need to filter phrase table or decoding with CN and Lattices not feasible

# Summary

- MT
  - Ran a lot of experiments!
  - Sequence models (LMs) over Brown clusters help
  - OSM models over Brown clusters generally help more than LMs
  - Combining OSM models over multiple representations (POS+Morph tags) helps en-de and de-en
- Take home message:

  OSM models over Brown clusters will help