

# Morphological Processing and Word Reordering for Statistical MT of Highly Inflected Languages

Marcello Federico Arianna Bisazza Christian Hardmeier  
Human Language Technologies Research Unit  
FBK-irst, Trento - Italy

Haifa, 24 January 2011

- Statistical MT in a nutshell
- When it works and when it does not
- Case study 1: Turkish to English
- Case study 2: Arabic to English
- Case study 3: German to English
- Conclusions

## To take home:

embedding morpho-syntactic information into SMT is possible and it works!

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the European Commission under the 7<sup>th</sup> Framework Programme for Research and Technological Development.

Freedom of movement must be encouraged , while ensuring that career paths are safeguarded .

E' necessario incoraggiare tale mobilità pur garantendo la sicurezza dei percorsi professionali .

Freedom of movement must be encouraged , while ensuring that career paths are safeguarded .

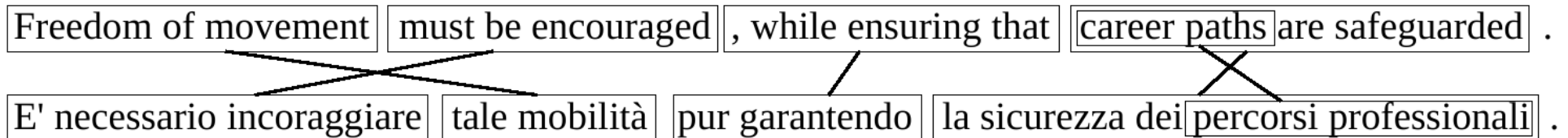
E' necessario incoraggiare tale mobilità pur garantendo la sicurezza dei percorsi professionali .

## How SMT works (in a nutshell)

- **operations**: segment, translate, and place
- **scores**: linear combination of feature functions
- **features**: phrase pairs, target n-grams, relative phrase movement , ...
- **search**: efficient algorithm to compute (sub-)optimal solutions
- features and combination weights are **machine learnable** from parallel data

Freedom of movement must be encouraged , while ensuring that career paths are safeguarded .

E' necessario incoraggiare tale mobilità pur garantendo la sicurezza dei percorsi professionali .



## When SMT works (when "more data" is not enough)

- **simple morphology** of source/target
  - better n-gram models, better alignments, less OOV words, ...
- **similar morphology** between source and target
  - better alignments, richer phrase tables, ...
- **similar syntax** between source and target
  - better alignments, phrase-tables, word re-ordering,...

For many language pairs we are far from the ideal condition.

What can we do? what has been done?

- **Enhance SMT features** to capture more information
  - factored models, shallow/deep syntax models, hierarchical re-ordering model
- **Integrate language knowledge** within the existing models
  - morphology pre-preprocessing, word-order pre-processing

We report recent work on the second approach for three translation directions:

- **Turkish** to English, IWSLT BTEC task
- **Arabic** to English, NIST MT 2009 task
- **German** to English, WMT 2010 task

All case studies are carried out with the Moses and IRSTLM toolkits.

## Morphological Pre-processing for Turkish SMT

A. Bisazza, M. Federico. “Morphological Pre-Processing for Turkish to English SMT.” Proc. of International Workshop on Spoken Language Translation, 2009.

A. Bisazza, I. Klasinas, M. Cettolo, M. Federico. “FBK @ IWSLT 2010.” Proc. of the International Workshop on Spoken Language Translation, 2010.

Tourist expressions: simple task but limited training data

Rich morphology of Turkish has negative impact on SMT

	Training		OOV on Test (iwslt04)
	W	V	
TR	139.5K	17.6K	6.7%
EN	182.6K	8.3K	

Examples:

SRC: Belki bir doktora görünmelisin.

REF: Perhaps you should see a doctor.

OUT: Maybe [*görünmelisin*] a doctor.

SRC: Bu film rulolarını banyo ettirip basabilir miydiniz?

REF: Could you develop and print these rolls of film?

OUT: Could you reissue [*ettirip*] [*rulolarını*] this film developed ?

Several linguistic features can negatively affect an SMT system:

- **Agglutination**

→ vocabulary built by a wide range of suffix combinations

<i>oda</i>		'room'
<i>odam</i>		' <b>my</b> room'
<i>odamda</i>		' <b>in</b> my room'
<i>odamdayım</i>		' <b>I am</b> in my room'

- **Vowel harmony** and other phoneme alternation phenomena

→ systematic stem and suffix *allomorphy*

<i>saç + (I)m</i>	→	<i>saç<u>ım</u></i>	'my hair'
<i>el + (I)m</i>	→	<i>el<u>im</u></i>	'my hand'
<i>kol + (I)m</i>	→	<i>kol<u>um</u></i>	'my arm'
<i>göz + (I)m</i>	→	<i>göz<u>üm</u></i>	'my eye'
<i>kafa + (I)m</i>	→	<i>kaf<u>a</u>m</i>	'my head'



**Idea:** selectively isolating or removing suffixes from the words

**Workflow:**

1. Morphological analysis and suffix normalization [Oflazer, 94]:  
suffix boundaries are detected and surface forms are replaced by tags to address vowel harmony and allomorphy.
2. Morphological disambiguation in context [Sak and Saraclar, 2007]:  
only the most likely analysis is taken for each word
3. Rules for splitting/removal of suffix tags:  
15 segmentation schemes developed and tested. Best performing schemes:
  - *MS11*: handles **nominal** suffixes (case, possessive) and copula;
  - *MS13*: also isolates **verbal** negation suffix;
  - *MS15*: also isolates other **verbal** suffixes: subject person, ability & voice.

Examples: surface form *vs* normalized representation:

*I was in my room*  
 = *odamdaydım* → *oda / m / da / ydı / m*  
 [room-my-in-was-I] [room] [my] [in] [was] [I]

*I can not explain*  
 = *anlatamıyorum* → *anla / t / a / mı / yor / um*  
 [understand-make-can-not-I] [understand] [make] [can] [not] [I]

Examples: surface form *vs* normalized representation:

*I was in my room*  
 = *odamdaydım* → *oda / m / da / ydı / m*  
 [room-my-in-was-I] [room] [my] [in] [was] [I]

---

oda+A3sg / +P1sg / +Loc / +Zero+Past / +A1sg  
 ↑ ↑ ↑ ↑ ↑  
 lemma poss. case copula person

*I can not explain*  
 = *anlatamıyorum* → *anla / t / a / mı / yor / um*  
 [understand-make-can-not-I] [understand] [make] [can] [not] [I]

---

anla+Prog1 / +Caus / +Able / +Neg / +A1sg  
 ↑ ↑ ↑ ↑ ↑  
 lemma+tense causative ability negation person

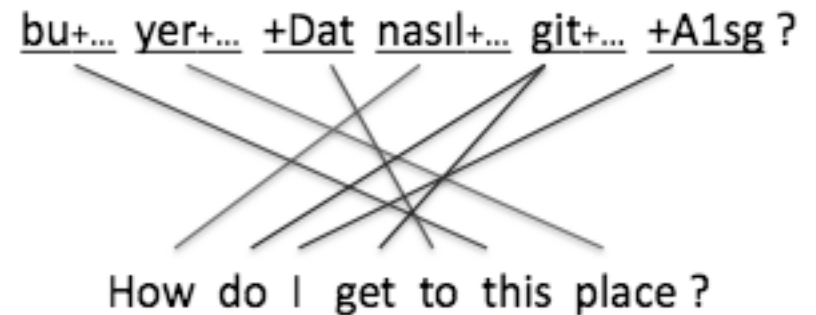
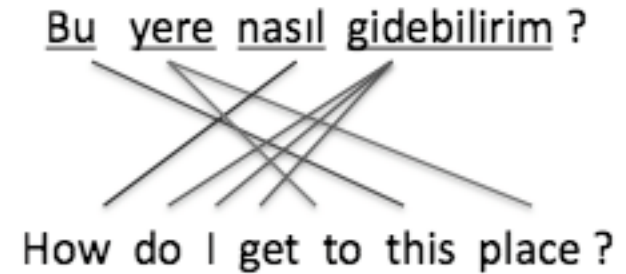
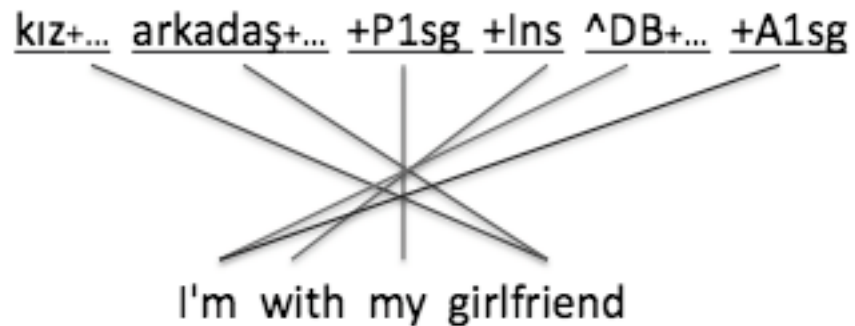
⇒ The underlying representation is used to train the SMT system.

## Results:

- minimizes differences in word granularity between TR and EN,
- abstracts from allomorphy by using a tag-like notation,
- reduces data sparseness, training dictionary size, OOV rate of test:

Preprocessing		Training		OOV on Test (iwslt04)
		$ W $	$ V $	
TR	basic tokenization	139.5K	17.6K	6.7%
	MS11	168.1K	10.4K	2.6%
	MS15	174.5K	9.5K	2.0%
<i>EN</i>	<i>basic tokenization</i>	<i>182.6K</i>	<i>8.3K</i>	–

- yields more refined alignments:



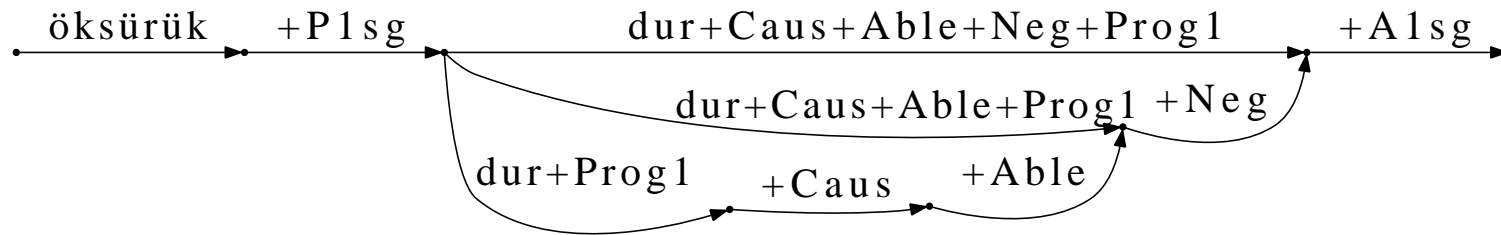
**Idea:** replace OOVs in the test by morphologically similar words seen in training:

- possible replacers: all words sharing the same lemma
- heuristic: choose candidates with tag sequence most similar to the OOV word
- OOVs replaced by  $n$ -best candidates in a confusion network input

Word	Gloss	Preprocessed (MS11)	Score <sup>1</sup>
<b><i>çıkışlar</i></b>	<b><i>exits</i></b>	<b>çık+Verb+Pos^DB+Noun+Inf3+A3pl</b>	
çıkış	exit	çık+Verb+Pos^DB+Noun+Inf3+A3sg	<b>93</b>
çıkma	going out	çık+Verb+Pos^DB+Noun+Inf2+A3sg	66
çıkacak	will go out	çık+Verb+Pos^DB+Noun+FutPart+A3sg	66
çıkan	who goes out	çık+Verb+Pos^DB+Adj+PresPart	44
çıkıyor	is going out	çık+Verb+Pos+Prog1	27
çıkıyor	isn't going out	çık+Verb+Neg+Prog1	0
çıkıyor	takes out	çık+Verb^DB+Verb+Caus+Pos+Aor	-15

<sup>1</sup>Score =  $20C - 2D_1 - 5D_2$ , where  $C$ : # of shared contiguous tags,  $D_1$ : # of different tags in the OOV,  $D_2$ : # of different tags in the candidate.

- Choice of optimal decomposition ruleset depends on task & target language
- Possible approach: combine various degrees of decomposition in input  
 ⇒ decoder can choose word-level-optimal segmentation path
- Training set = differently segmented versions of train, concatenated
- Example lattice combining MS11 + MS13 + MS15:



TR: öksürüğümü durduramıyorum

(EN: I cannot make my cough stop)

System \ BLEU	iwslt04	iwslt09	iwslt10
baseline	54.80	–	–
segm. ruleset MS11	60.30	57.21	52.14
segm. ruleset MS15	60.32	<b>58.28</b>	52.46
MS11 + lexical approx.(3-best)	59.68	57.11	51.76
segm. lattice MS11+13+15	<b>60.41</b>	57.70	<b>53.29</b>

- Morphological decomposition yields substantial improvements on baseline
- Adding rules for verbal inflection (MS15) helps slightly but consistently
- Lexical approximation unfortunately doesn't help
- Decomposition lattice works best for two of the three test sets

**Conclusions:** choice of pre-processing technique depends on task and dataset.



## Morphological Pre-processing for Arabic SMT

N. Bertoldi, A. Bisazza, M. Cettolo, M. Federico and G. Sanchis-Trilles. “FBK @ IWSLT 2009”. Proc. of the International Workshop on Spoken Language Translation, 2009.

Rich morphology, but also orthographic variations and different vowelization styles.  
→ specific preprocessing reduces data sparseness and improves alignments.

**Arabic tokenization:** Unicode characters and digits normalization, removal of diacritics and *tatweel* (justification character).

**Morphological decomposition:** isolates clitics from words.

Two state-of-the-art linguistic tools compared:

- **MADA**
  - heavy-weight: based on linguistic features produced by Buckwalter analyzer,
  - optimised use of the tool to run on large corpora
- **AMIRA**
  - light-weight: SVM classifier based on a  $-5/+5$  character context.

Two different segmentation schemes:

- MADA (scheme D2) splits prefixes: conjunctions (w+ 'and', f+ 'then'), prepositions (b+ 'by', k+ 'as', l+ 'to'), future tense mark (s+).  
Also normalizes orthography (beginning *alef*, *tah marbutah*, *alef maksura* . . .)
- AMIRA doesn't split future mark, but splits suffixes: object and poss. pronouns.

	<i>'and she will say it to her colleague':</i>						
<b>Baseline</b>	wstqwlh			lzmylhA			
	[and-she-will-say-it]			[to-her-colleague]			
<b>MADA</b>	w+	s+	tqwlh	l+	zmylhA		
	[and]	[will]	[she-say-it]	[to]	[her-colleague]		
<b>AMIRA</b>	w+	stqwl	+h	l+	zmyl	+hA	
	[and]	[she-will-say]	[it]	[to]	[colleague]	[her]	

On the NIST task MADA slightly outperforms AMIRA, but AMIRA is faster and includes shallow chunking.

## Verb Reordering for Arabic SMT

A. Bisazza and M. Federico. “Chunk-based verb reordering in VSO sentences for Arabic-English SMT.” Proc. of ACL Workshop on SMT and Metrics, 2010.

A. Bisazza, D. Pighin, M. Federico. “Chunk-Lattices for Verb Reordering in Arabic-English SMT”. *Machine Translation*, 2010. (Accepted for publication).

## Problem:

Word reordering is a challenge for phrase-based SMT between distant languages

English: mainly Subject-Verb-Object VS Arabic: both SVO and VSO

Typical errors in phrase-based SMT outputs:

\*The Moroccan monarch King Mohamed VI  $\emptyset$  his support to the French President

\*He renewed the Moroccan monarch King Mohamed VI his support to the French President

- Focus on verbs *say, declare, note...* in pre-subject position of news
- Apply simple surface pattern-matching reordering rules, without syntax
- Rule: move **verb** before **trigger element** (*'that', colon, quotation mark, etc.*)

## Example 1

### original

src: qAlt hh AlwkAlp : nZrA l+ AlwDE AlHAly fy AlErAq ...  
 mt: She said the agency: In view of the current situation in Iraq ...

### reordered

src: h\*h AlwkAlp qAlt : nZrA l+ AlwDE AlHAly fy AlErAq ...  
 mt: The agency said due to the current situation in Iraq ...

- Focus on verbs *say, declare, note...* in pre-subject position of news
- Apply simple surface pattern-matching reordering rules, without syntax
- Rule: move **verb** before **trigger element** (*'that', colon, quotation mark, etc.*)

## Example 2

### original

src: tAbE byAn SAdr En mktb hnyp >n Al>xyr ...

mt: He went on to say, a statement issued by the office of Hania that the latter

### reordered

src: byAn SAdr En mktb hnyp tAbE >n Al>xyr ...

mt: A statement issued by the office of Hania continued that the latter ...

- Focus on verbs *say, declare, note...* in pre-subject position of news
- Apply simple surface pattern-matching reordering rules, without syntax
- Rule: move **verb** before **trigger element** (*'that', colon, quotation mark, etc.*)

## Example 2

### original

src: tAbE byAn SAdr En mktb hnyp >n Al>xyr ...

mt: He went on to say, a statement issued by the office of Hania that the latter

### reordered

src: byAn SAdr En mktb hnyp tAbE >n Al>xyr ...

mt: A statement issued by the office of Hania continued that the latter ...

Unfortunately, no significant BLEU improvement on the NIST task.

We introduce more linguistic knowledge and extend to all verbs!

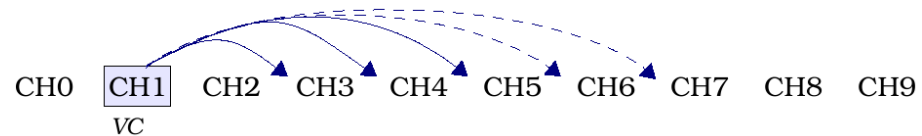


## Assumptions:

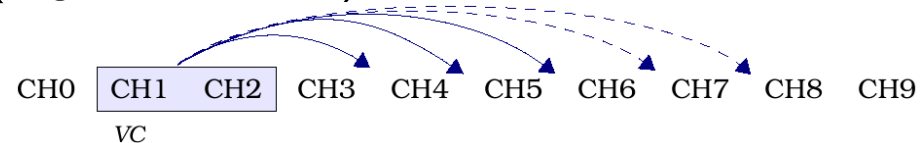
- 1) verb reordering only between shallow syntax chunks
- 2) no overlap between consecutive verb movements

Define a class of possible **movements**:

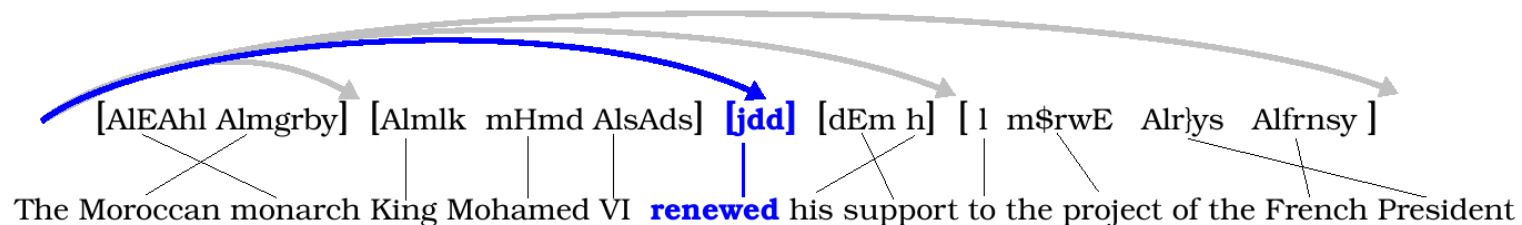
i) move verb chunk...



ii) ... or verb chunk + next chunk (e.g. adverbials)  
by up to N chunks to the right



Best movement in parallel corpus:  
minimizes global distortion wrt to English translation



The reordered parallel corpus is used to train the SMT system.  
As for the test, we use word **reordering lattices**.

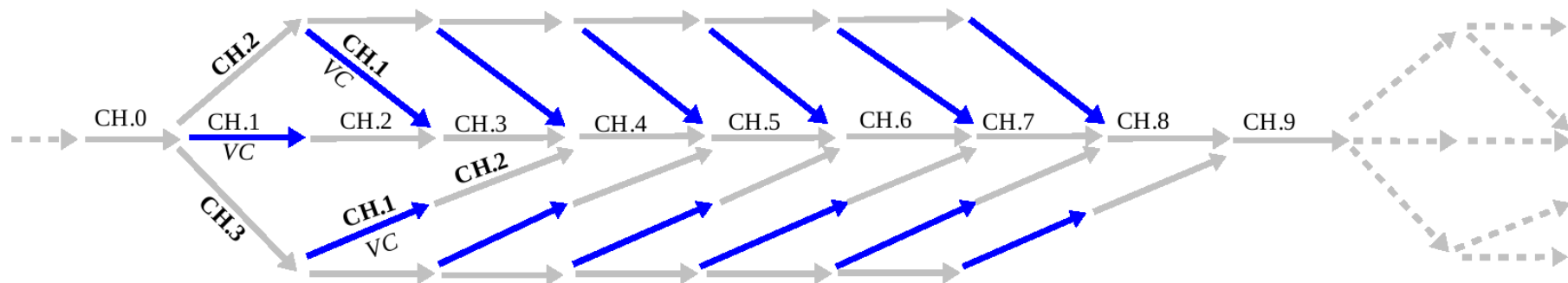
Given the initial assumptions, we can build compact lattices and run non-monotonic decoding on them (Dyer & al. 2008)

Hybrid approach:

- for verb reordering: lattices
- for other reorderings: standard (phrase-internal and local distortion)

Lattice representation of the rule:

“move 1 or 2 chunks by up to 6 chunk positions right”



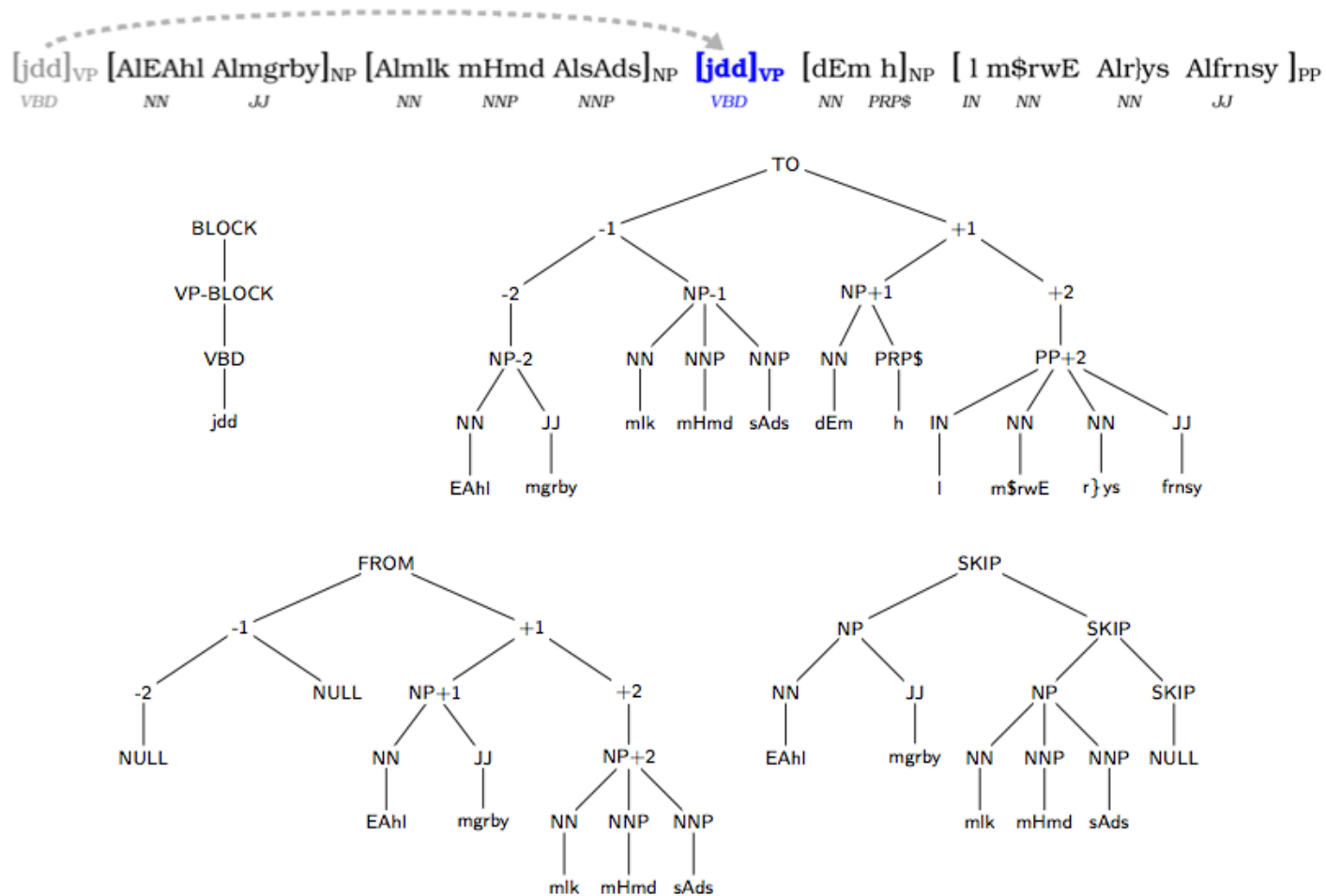
High-end baseline: Moses, 30M words newswire from NIST09  
with lexicalized reordering models (Och &al. 2004; Koehn &al. 2007)

Different experimental conditions:

- whole system re-trained and tuned on verb-reordered data
- translation of plain input (text)
- translation of reordering lattice

System	DL	Eval08-NW		Eval09-NW	
		bleu	krs <sup>2</sup>	bleu	krs
baseline	6	43.10	80.57	48.13	83.17
<hr/>					
reord. training +					
plain input	6	43.67	80.62	48.53	83.58
lattice	4	<b>44.04</b>	<b>80.93</b>	<b>48.96</b>	<b>83.75</b>
<hr/>					
<i>oracle reordered</i>	<i>4</i>	<i>44.36</i>	<i>81.29</i>	<i>49.26</i>	<i>84.30</i>

<sup>1</sup>Kendall Reordering Score: similarity btw word order of outputs and of references (Birch &al.2010)



We use *syntactic tree kernel* to represent verb chunk movements; Fig. shows forest corresponding to one specific movement. We train a SVM by optimizing global distortion in the training data.

System	DL	Eval08-NW		Eval09-NW	
		bleu	krs	bleu	krs
baseline	6	43.10	80.57	48.13	83.17
reord. training +					
full lattice	4	44.04	80.93	48.96	83.75
1-best-pruned	4	<b>44.34</b>	81.18	49.10	<b>84.15</b>
2-best-pruned	4	44.29	<b>81.30</b>	<b>49.19</b>	84.02
3-best-pruned	4	44.11	81.13	49.05	83.90

- Simply reordering of the training data is beneficial:  
more monotone alignments  $\Rightarrow$  better phrase extraction
- Providing likely reordering in the lattice yields further improvement
- Interesting: reordering-specific metric correlates well with BLEU
- Further improvement:
  - pruning the lattice with discriminative approach (SVM)

## Morphological Reduction and Reordering for German

C. Hardmeier, A. Bisazza and M. Federico. “FBK at WMT 2010: word lattices for morphological reduction and chunk-based reordering.” Proc. of ACL Workshop on SMT and Metrics, 2010.

## Morphology

- Inflectional morphology: much more prolific in German  
Nouns have case, verbs have many forms, etc.
- Derivational morphology:  
German has **one-word-compounds** that must be split  
→ many vocabulary types, high OOV rate

## Word order

- English: strict **SVO** word order
- German: **SVO** in main clauses, **SOV** in subordinate clauses  
→ word order mismatch

**Approach:** morphological reduction and chunk-based reordering

- We use **Gertwol** to split compounds and reduce words to their base form.  
Gertwol: commercial two-level finite-state morphology
- Gertwol analyses are disambiguated with POS tags and heuristic disambiguation rules (courtesy of the University of Zurich).
- **Decoding**: supply reduced forms as alternative paths in a lattice:



- **Training**: concatenate original and processed parallel texts.

	BLEU	
	DEV	EVAL
Baseline	18.8	20.1
with morphological reduction	19.3	20.6



- Same mechanism as for Arabic-English, but different rules.
- We concentrate on a few patterns involving verbs.
- Simplifying assumption:  
Verb reordering only occurs between shallow syntax chunks.
- Tagging and chunking done with the TreeTagger.
- Small number of hand-written reordering rules that can generate multiple reorderings for each matching verb chunk.

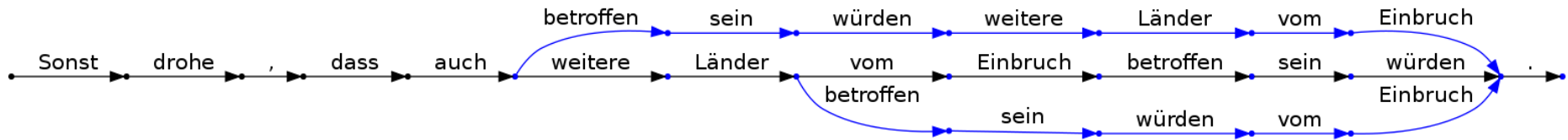
## Example: Subordinate clause rule

*Motivation* Move clause-final verbs in German **SOV** subordinates left to match English **SVO** word order.

*Moving block* Verb chunk immediately followed by punctuation.

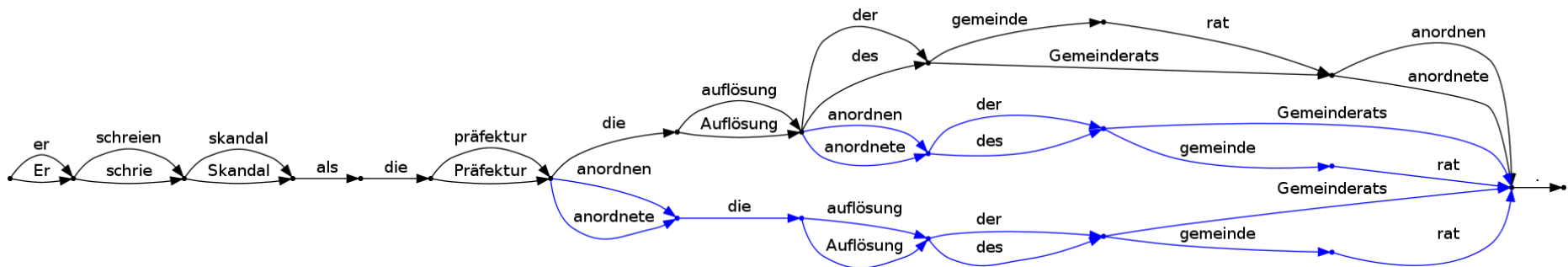
*Movement* to the left  
1 to 3 chunks after most recent subordinating conjunction

Sonst [drohe]<sub>VC</sub> , dass auch [weitere Länder]<sub>NC</sub> [vom Einbruch]<sub>PC</sub> [betroffen sein würden]<sub>VC</sub> .



It is straightforward to merge a morphological reduction lattice with a chunk reordering lattice:

[Er]<sub>NC</sub> [schrie]<sub>VC</sub> [Skandal]<sub>NC</sub> als [die Präfektur]<sub>NC</sub> [die Auflösung]<sub>NC</sub> [des Gemeinderats]<sub>NC</sub> [anordnete]<sub>VC</sub> .



# English-German: Results

	BLEU			
	DEV		EVAL	
	–	MR	–	MR
Baseline	18.8	19.3	20.1	20.6
with reordering	18.9	19.8	20.3	21.1

*MR = morphological reduction*

- Chunk reordering on its own helps very little: around 0.2 BLEU points.
- In combination with morphological reduction, the gain is much greater: half a point for morphological reduction + half a point for reordering = one point total improvement
- Reordering with lattices strongly depends on the language model to choose the right path.

- We showed methods to exploit **morpho-syntactic information** for SMT
  - that also resulted in **performance improvements** on strong baselines
- **Language expertise** of the source/target languages definitely helps
  - to identify, analyze, and describe issues from a linguistic perspective
- **Statistical modeling expertise** is required
  - to conceive, implement, and integrate new features in the decoder
  - to exploit or extend existing features
- The phrase-based SMT framework is **simple, flexible, and extensible**
  - there are more and more things that can be explored, improved, integrated
- **Current evolution** of the presented approaches:
  - re-ordering models embedding language specific syntactic constraints/preferences
  - context models to enforce cohesive MT across different sentences

- We showed methods to exploit **morpho-syntactic information** for SMT
  - that also resulted in **performance improvements** on strong baselines
- **Language expertise** of the source/target languages definitely helps
  - to identify, analyze, and describe issues from a linguistic perspective
- **Statistical modeling expertise** is required
  - to conceive, implement, and integrate new features in the decoder
  - to exploit or extend existing features
- The phrase-based SMT framework is **simple, flexible, and extensible**
  - there are more and more things that can be explored, improved, integrated
- **Current evolution** of the presented approaches:
  - integrate language-specific word-order knowledge directly in the decoder
  - embed syntactic knowledge in re-ordering models and future cost estimation

# thank you