A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena

Arianna Bisazza[*]
University of Amsterdam

Marcello Federico[**]
Fondazione Bruno Kessler

Word reordering is one of the most difficult aspects of statistical machine translation (SMT), and an important factor of its quality and efficiency. Despite the vast amount of research published to date, the interest of the community in this problem has not decreased, and no single method appears to be strongly dominant across language pairs. Instead, the choice of the optimal approach for a new translation task still seems to be mostly driven by empirical trials.

To orient the reader in this vast and complex research area, we present a comprehensive survey of word reordering viewed as a statistical modeling challenge and as a natural language phenomenon. The survey describes in detail how word reordering is modeled within different string-based and tree-based SMT frameworks and as a stand-alone task, including systematic overviews of the literature in advanced reordering modeling.

We then question why some approaches are more successful than others in different language pairs. We argue that besides measuring the amount of reordering, it is important to understand which kinds of reordering occur in a given language pair. To this end, we conduct a qualitative analysis of word reordering phenomena in a diverse sample of language pairs, based on a large collection of linguistic knowledge. Empirical results in the SMT literature are shown to support the hypothesis that a few linguistic facts can be very useful to anticipate the reordering characteristics of a language pair and to select the SMT framework that best suits them.

1. Introduction

Statistical machine translation (SMT) is a data-driven approach to the translation of text from one natural language into another. It emerged in the 1990s and matured in the 2000s to become widespread today; the core SMT methods (Brown et al. 1990, 1993; Berger et al. 1996; Koehn, Och, and Marcu 2003) learn direct correspondences between source and target language from collections of translated sentences, without the need for

[*] Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. E-mail: a.bisazza@uva.nl.

[**] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Trento, Italy. E-mail: federico@fbk.eu.

Submission received: 5 February 2015; revised version received: 3 November 2015; accepted for publication: 12 January 2016.

doi:10.1162/COLI_a_00245

	جدد العاهل المغربي الملك محمد السادس دعم ـه لـ مشروع الرئيس الفرنسي											
SRC	*verb*	*subject*					*object*		*complement*			
	jdd	AlEAhl	Almgrby	Almlk	mHmd	AlsAds	dEm	-h	l-	m$rwE	Alr}ys	Alfrnsy
	renewed	*the-monarch*	*the-Moroccan*	*the-King*	*Mohamed*	*the-sixth*	*support*	*his*	*to*	*project*	*the-president*	*the-French*
REF	The Moroccan monarch King Mohamed VI **renewed** his support to the project of the French President											

Figure 1
Arabic source sentence (right-to-left) and English reference translation, taken from the NIST-MT09 benchmark. The Arabic sentence is morphologically segmented by AMIRA (Diab, Hacioglu, and Jurafsky 2004) according to the Arabic Treebank scheme, and provided with Buckwalter transliteration (left-to-right) and English glosses.

abstract linguistic representations. The main advantages of SMT are versatility and cost-effectiveness: In principle, the same modeling framework can be applied to any pair of languages with minimal engineering effort, given sufficient amounts of translation data. However, experience in a diverse range of language pairs has revealed that this form of modeling is highly sensitive to structural differences between source and target language, particularly at the level of word order.

Indeed, natural languages vary greatly in how they arrange sentence components, and translating words in the correct order is essential to preserving meaning across languages. In English, for instance, the role of different predicate arguments is determined precisely by their relative position within the sentence. Consider the translation example in Figure 1: Looking at the English glosses of the Arabic sentence, one can see that corresponding words in the two languages are placed in overall similar orders with the notable exception of the verb (*jdd/renewed*), which occurs at the beginning of the Arabic sentence but in the middle of the English one—more specifically, between the subject and the object. To reach the correct English order, three other reorderings are required between pairs of adjacent Arabic words: (*AlEAhl/the-monarch*, *Almgrby/the-Moroccan*), (*dEm/support*, *-h/his*), and (*Alr}ys/the-president*, *Alfrnsy/the-French*). This example suggests a simple division of reordering patterns into long range, or global, and short range, or local. However, other language pairs display more complex, hierarchical patterns.

Word reordering phenomena are naturally handled by human translators[1] but are a major source of complexity for SMT. In very general terms, the task of SMT consists of breaking the input sentence into smaller units, selecting an optimal translation for each unit, and placing them in the correct order. Searching for the overall best translation throughout the space of all possible reorderings is, however, computationally intractable (Knight 1999). This crucial fact has motivated an impressive amount of research around two inter-related questions: namely, how to effectively restrict the set of allowed word permutations and how to detect the best permutation among them.

Existing solutions to these problems range from heuristic constraints, based on word-to-word distances and completely agnostic about the sentence content, to linguistically motivated SMT frameworks where the entire translation process is guided by syntactic structure. The research in word reordering has advanced together with core SMT research and has sometimes directed it, being one of the main motivations for the development of tree-based SMT. At the same time, the variety of word orders existing in world languages has pressed the SMT community to admit the importance of

[1] Nevertheless, learning and understanding a new language has been shown to be more difficult when the new language is structurally distant from one's native language (Corder 1979).

language-specific knowledge and to reassess its ambitions towards a universal translation algorithm.

According to the Machine Translation Archive, a scientific interest in this specific subproblem of MT started around 2006 and kept growing at a rapid pace. In 2014, the research papers mainly dedicated to reordering accounted for no less than 10% of all SMT papers.[2] Despite the abundant research, word order differences remain among the most important factors of performance in modern SMT systems, and new approaches to reordering are still proposed every year.

To orient the reader in this complex and productive research area, we present a comprehensive survey of word reordering viewed as a statistical modeling challenge and as a natural language phenomenon. Our survey notably differs from previous work (Costa-jussà and Fonollosa 2009) in that we not only review the existing approaches to word reordering in SMT, but we also question why some approaches are more successful than others in different language pairs. In particular, we argue that understanding the complexity of reordering in a given language pair is key to selecting the right SMT models and to improving them.

The survey is organized as follows: Section 2 explains how the word reordering problem is treated within different string-based and tree-based SMT frameworks, as well as a stand-alone task (i.e., pre- and post-ordering). The literature in advanced reordering modeling is extensively reviewed, with a major focus on recent work. Section 3 describes the challenges of automatically assessing word reordering accuracy in SMT outputs. Section 4 presents a qualitative analysis of word reordering across language pairs. In particular, detailed word order profiles are provided for a sample of seven widely spoken languages representing structural and geographical diversity: namely, English, German, French, Arabic, Turkish, Japanese, and Chinese. The same section reviews empirical results from the SMT literature, showing that the proposed word order profiles are useful to anticipate the reordering characteristics of a language pair and to select the SMT framework that best suits them. The survey ends with a discussion of the strengths and weaknesses of the major approaches to reordering in SMT.

2. Approaches to Word Reordering in Statistical Machine Translation

A first important distinction has to be made between word reordering performed as part of the decoding process (Sections 2.1 to 2.3) and word reordering performed *before* or *after* it as a monolingual task decoupled from the bilingual translation task (Section 2.4).

Within the former, we further distinguish between string-based (sequential) approaches and tree-based (structural) approaches. **String-based SMT** (Sections 2.1 and 2.2) treats translation as a **sequential** task: The target sentence is built from left to right while the input units are visited in different orders and no dependencies other than word adjacency are considered. Subsequently, problem decomposition is applied to the target **string**: an optimal translation is sought for each prefix of the target translation, from the shortest to the longest. **Tree-based SMT** (Section 2.3) posits the existence of a **tree structure** to explain translation as a hierarchical process and to capture dependencies among non-adjacent text units. Problem decomposition is therefore based on this structure: An optimal translation is sought for each word span corresponding to a node in the tree, from the leaves up to the root. Whereas string-based SMT has to search over

[2] Peer-reviewed conferences, workshops, and journal papers listed by the Machine Translation Archive: `http://www.mt-archive.info/srch/subjects.htm`.

all input permutations that do not violate some general reordering constraints, tree-based SMT considers only those permutations that result from transforming a given tree representing the input sentence (as for example permuting each node's children).

Moreover, we should note the difference between **syntax-based SMT** approaches that utilize trees produced by monolingual parsers trained on syntactic treebanks and **data-driven tree-based SMT** approaches that extract bilingual translation grammars directly from pairs of source and target sentences. In the former, word reordering is constrained by a given syntactic parse tree of the input sentence or by the grammar of the target language (or both), whereas in the latter, tree structure captures hierarchical reordering patterns that may or may not correspond to syntactically motivated rules.

In general, the SMT search (or decoding) process consists in searching for the most probable target (or English) sentence $\mathbf{e}^*$ given a source (or foreign) sentence $\mathbf{f}$ by scoring translation hypotheses through a linear combination of feature functions:

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} \max_{\mathbf{b}} \ \exp\left[\sum_{r=1}^{R} \lambda_r \mathrm{h}_r(\mathbf{f}, \mathbf{e}, \mathbf{b})\right] \tag{1}$$

where $\mathbf{b}$ is a latent variable representing either a linear or a hierarchical mapping (alignment) between $\mathbf{f}$ and $\mathbf{e}$, $\mathrm{h}_r(\mathbf{e}, \mathbf{f}, \mathbf{b})$ are R arbitrary feature functions and λ_r the corresponding feature weights. Feature functions try to capture relevant translation adequacy and word reordering aspects from aligned parallel data, as well as translation fluency aspects from monolingual target texts. Moreover, feature functions are assumed to be locally decomposable to allow for efficient decoding via dynamic programming. Feature weights are tuned discriminatively by directly optimizing translation quality[3] on a development set, using parameter tuning techniques such as MERT (Och 2003), MIRA (Chiang, Marton, and Resnik 2008), or PRO (Hopkins and May 2011).

2.1 Phrase-Based SMT

Phrase-based SMT (PSMT) is the currently dominant approach in string-based SMT. PSMT ruled out the early word-based SMT framework (Brown et al. 1990, 1993; Berger et al. 1996) thanks to two important novelties: the use of multi-word translation units (Och 1999; Zens, Och, and Ney 2002; Koehn, Och, and Marcu 2003), and the move from a generative to a discriminative modeling framework (Och and Ney 2002).

The search process (1) in PSMT is guided by the target string $\mathbf{e}$, built from left to right, and the alignment variable $\mathbf{b}$ that embeds both segmentation and reordering of the source phrases. This is defined as

$$\mathbf{b} = b_1^I = ((J_1, K_1), (J_2, K_2), \ldots, (J_I, K_I)) \tag{2}$$

such that $K_1, \ldots, K_I$ are consecutive intervals partitioning the target word positions, and $J_1, \ldots, J_I$ are corresponding but not necessarily consecutive intervals partitioning the source word positions. A phrase segmentation for our running example is shown in Figure 2.

The use of phrases mainly results in a better handling of ambiguous words and many-to-many word equivalences, but it also makes it possible to capture a considerable amount of local reordering phenomena within a translation unit (*intra*-phrase

[3] Automatic measures of translation quality are discussed in Section 3.

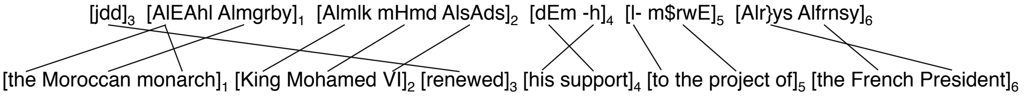

[jdd]3 [AlEAhl Almgrby]1 [Almlk mHmd AlsAds]2 [dEm -h]4 [l- m$rwE]5 [Alr}ys Alfrnsy]6
[the Moroccan monarch]1 [King Mohamed VI]2 [renewed]3 [his support]4 [to the project of]5 [the French President]6

Figure 2
An example of word alignment and phrase segmentation for the sentence pair presented in Figure 1. Subscript indices denote the phrase alignment b_1^I. Note that other phrase segmentations are possible given the same word alignment.

reordering). With reference to our running example (Figure 1), a PSMT model may handle the local reorderings as single phrase pairs—*[AlEahl Almgrby]-[The Moroccan monarch]*, and so forth—if these were observed in the training data. On the contrary, it is unlikely that a single long phrase spanning from *jdd* to *AlsAds* was observed, therefore the long-range reordering of the verb could be handled by *inter*-phrase reordering.

State-of-the art PSMT systems typically include the following core feature functions: phrase- and word-level translation models; target *n*-gram language model; distortion penalty; plus additional components that model specific translation aspects. Assuming a one-to-one correspondence between source and target phrases, reordering in PSMT means searching through a set of permutations of the source phrases. Thus, two subproblems arise: defining the set of permutations in **b** allowed during decoding (reordering constraints) and scoring the allowed permutations (reordering models or feature functions). We will now discuss each of them in detail.

2.1.1 PSMT Reordering Constraints. Because searching over the space of all possible translations is NP-hard (Knight 1999), SMT decoders use heuristic search algorithms to only explore a promising subset of the search space. In particular, limiting the set of explorable input permutations is an essential way to reduce decoding complexity.

The reordering constraint originally included in the PSMT framework is called the **distortion limit (DL)**. This consists in allowing the decoder to jump, or skip, at most *k* words between the last translated source phrase and the next one, that is:

$$jump(J_{i-1}, J_i) = |start(J_i) - end(J_{i-1}) - 1| \ \leq \ k \tag{3}$$

Setting a low distortion limit means only exploring local reorderings, based on the arguable assumption that languages tend to arrange sentence constituents in similar orders. Besides being essential for efficiency—DL allows for linear decoding complexity—reordering constraints are also important for translation quality because the existing SMT models are typically not discriminative enough to guide the search over very large sets of reordering hypotheses. However, reordering constraints have also several drawbacks. For instance, the verb reordering in Figure 2 may not be captured by a PSMT system that applies a DL of $k = 5$ or less, because jumping back from *AlsAds* to *jdd* corresponds to a skip of six positions. While the distortion limit is a de facto standard in modern PSMT systems, the first constraining paradigms were formulated earlier for word-based SMT (Berger et al. 1996; Zens and Ney 2003) and are called IBM constraints.

A different kind of reordering constraint can be derived from the Inversion Transduction Grammars (**ITGs**) (Wu 1995, 1997). **ITG constraints** only admit permutations that are generated by recursively swapping pairs of adjacent blocks of words.[4] In particular, ITG constraints disallow reorderings that generalize the patterns (3 1 4 2) and (2 4 1 3), which are rarely attested in natural languages (Wu 1997).[5] Enforcing ITG constraints in left-to-right PSMT decoding requires the use of a shift-reduce permutation parser (Zens 2008; Feng et al. 2010). Alternatively, a relaxed version of the ITG constraints (i.e., Baxter permutations) may be enforced by simply inspecting the set of covered source positions, as proposed by Zens et al. (2004) and Zens (2008). Interestingly, Cherry, Moore, and Quirk (2012) found no consistent benefit from applying either exact or approximate ITG-constraints to a PSMT system that already included a hierarchical phrase orientation model[6] (Galley and Manning 2008).

The reordering constraints presented so far are not sensitive to the words being translated nor to their context. This results in a very coarse definition of the reordering search space, which is problematic in language pairs with different syntactic structures. To address this problem, Yahyaei and Monz (2009) propose decoupling local and global reordering by segmenting the input sentence into chunks that can be permuted arbitrarily, but each of which is translated monotonically. In related work, Yahyaei and Monz (2010) present a technique to *dynamically set the DL* during decoding: They train a discriminative classifier to predict the most probable jump length after each input word, and use the predicted value as the DL after that position. Unfortunately, this method appears to generate inconsistent constraints leading to decoding dead-ends. Bisazza and Federico (2013a) further develop this idea so that only long reorderings predicted by a specific reordering model are explored by the decoder. This form of early reordering pruning enables the PSMT system to capture long-range reordering without hurting efficiency and is not affected by the constraint inconsistency problem.

When available, a parse tree of the input may also be used to constrain PSMT reordering, following the principle of **syntactic cohesion** (Fox 2002). Concretely, the dependency cohesion constraint (Cherry 2008) states that when part of a source subtree is translated, all words under the same subtree must be covered before moving to words outside of it. Integrated in phrase-based decoding as soft constraints (i. e., by using the number of violations as a feature function), dependency cohesion and its variants (Cherry 2008; Bach, Vogel, and Cherry 2009) were shown to significantly improve translation quality. In related work, Feng, Sun, and Ney (2012) derive similar cohesion constraints from the semantic role labeling structure of the input sentence. The divide-and-translate approach of Sudoh et al. (2010) uses source-side parse trees to segment complex sentences into simple clauses which are replaced by specific symbols and translated independently. Then, the target sentence is reconstructed using the placeholders, with the aim of simplifying long-range clause-level reordering.

2.1.2 PSMT Reordering Feature Functions. Target language modeling is the primary way to reward promising reorderings during translation. This happens indirectly, through the scoring of target word *n*-grams, which are generated by translating the source positions in different orders. However, the fixed-size context of language models used in SMT (typically four or five words) makes them largely insensitive to global reordering phenomena. In recent years, a growing interest in language pairs with very different word orders, such as Arabic–English and Chinese–English, has favored the

[4] For a comparative study of the IBM and ITG constraints, we refer the reader to Zens and Ney (2003).
[5] Empirical evidence against this was presented by Wellington, Waxmonsky, and Melamed (2006).
[6] The reordering models mentioned herein are explained in detail in the next section.

development of new techniques to explicitly model the reordering problem. Given a source sentence, the search for its optimal reordering is generally decomposed into a sequence of local reordering decisions, as is done for the whole translation process. Thus, the basic reordering step corresponds to the relative positioning of the word or phrase being translated, with respect to the word or phrase that was previously translated.

The simplest example of reordering feature function is the **distortion cost** or distortion penalty *jump*(J_{i-1}, J_i), which by convention assigns zero cost to hypotheses that preserve the order of the source phrases (monotonic translations). During decoding, the basic implementation of distortion cost penalizes long jumps only when they are performed, leading to the proliferation of hypotheses with gaps (i.e., uncovered input positions). This issue can be addressed by incorporating into the distortion cost an estimate of the cost yet to be incurred (Moore and Quirk 2007).

State-of-the-art systems use the distortion cost in combination with more sophisticated reordering models that take into account the identity of the reordered phrases and, optionally, various kinds of contextual information. A representative selection of such models is summarized in Table 1. To ease the presentation, we have divided the models into four groups according to their problem formulation: phrase orientation models, jump models, source decoding sequence models, and operation sequence models.

Phrase orientation models (POM) (Tillmann 2004; Koehn et al. 2005; Nagata et al. 2006; Zens and Ney 2006; Li et al. 2014), simply known as lexicalized reordering models, predict whether the next translated source span should be immediately to the right (monotone), immediately to the left (swap), or anywhere else (discontinuous) relatively to the last translated one.[7] For example, in Figure 2, the phrase pair *[Almlk mHmd AlsAds]-[King Mohamed VI]* has monotone orientation whereas *[jdd]-[renewed]* has discontinuous left orientation with respect to the previously translated phrase. Because of their simple reordering step classification, POM can be conditioned on very fine-grained information, such as the whole phrase pair, without suffering too much from data sparseness. However, because POM ignore the distance between consecutively translated phrases, they cannot properly handle long-range reordering phenomena and are typically used with a low distortion limit.

Jump models (JM) (Al-Onaizan and Papineni 2006; Green, Galley, and Manning 2010) predict the direction and length of the jump that is performed between consecutively translated words or phrases, with the goal of better handling long-range reordering. Because of data sparseness, JM work best when trained in a discriminative fashion using a variety of binary features (such as the last translated word, its POS tag, and relative position in the sentence) and when length bins are used instead of the exact jump length (Green, Galley, and Manning 2010). A drawback of JM is that they typically over-penalize long jumps because they are more rarely observed than short jumps.

Source decoding sequence models (SDSM) address this issue by directly modeling the reordered sequence of input words, as opposed to the reordering operations that generated it. This in turn can be done in several ways, such as: training n-gram models on target-like reordered source sentences and using them to score the sequence of input words visited by the decoder (Feng, Mauser, and Ney 2010); tagging the whole input sentence with symbols denoting how each word should be reordered with respect to its left and right context, then rewarding the decoding paths that most agree with the tag sequence (Feng, Peter, and Ney 2013); and finally, predicting which input position is

[7] Some phrase orientation models further distinguish between discontinuous left and discontinuous right.

Table 1
An overview of state-of-the-art reordering models for PSMT. Model type indicates whether a model is trained in a generative (gener.) or discriminative (discr.) way. All examples refer to the sentence pair shown in Figure 2.

Reordering models	References	Model type	Reordering step classification	Features
Phrase orientation models (POM): Example: $P($*orient=discontinuous-left* | *next-phrase-pair=[jdd]-[renewed]*$)$				

lexicalized (hierarchical) phrase orientation model	Tillmann 2004;
	Koehn et al. 2005;
	Nagata et al. 2006;
	Galley & Manning 2008
		gener.	monotonic, swap, discontinuous (left or right)	source/target phrases

phrase orientation maxent classifier	Zens & Ney 2006	discr.		source/target words or word clusters
sparse phrase orientation features	Cherry 2013	discr.		
Jump models (JM): Example: $P($*jump*$=-5$ | *from=AlsAds*, *to=jdd* $)$				
inbound/outbound/pairwise lexicalized distortion	Al-Onaizan & Papineni 2006	gener.	jump length	source words
inbound/outbound length-bin classifier	Green et al. 2010	discr.	jump length based (9 length bins)	source words, POS, position; sent. length
Source decoding sequence models (SDSM): Example: $P($*next-word=jdd* | *prev-translated-words=AlEahil Almlk mHmd AlsAds*$)$				
reordered source n-gram	Feng et al. 2010a	gener.	—	source words (9-gram context)
source word-after-word	Bisazza & Federico 2013; Goto et al. 2013	discr.	—	source words, POS; source context's words and POS
Operation sequence models (OSM): Example: $P(next - operation = generate[jdd,renewed]$ | *prev-operations=generate[AlsAds,VI] jumpBack[1]*$)$				

translation/reordering operation n-gram	Durrani et al. 2011;
	Durrani et al. 2013;
	Durrani et al. 2014
		gener.	insertGap,
			jumpBack,
			jumpForward
				source/target words, POS or word clusters; prev. n –1 operations

likely to be translated right after a given input position by means of a maximum entropy model using word and context features (Bisazza and Federico 2013a; Goto et al. 2013).

Operation sequence models (OSM) (Durrani, Schmid, and Fraser 2011) are n-gram models that include lexical translation operations and reordering operations (*insertGap, jumpBack,* or *jumpForward*) in a single generative story, thereby combining elements from the previous three model families. An operation sequence example is provided in the lower part of Table 1. OSM are closely related to n-gram based SMT models (see next section) but have been successfully applied as feature functions to PSMT (Durrani et al. 2013). To overcome data sparseness, OSM can be successfully applied to POS-tags and unsupervised word clusters (Durrani et al. 2014).

SDSM and OSM have been proven optimal for language pairs where high distortion limits are required to capture long-range reordering phenomena (Durrani, Schmid, and Fraser 2011; Bisazza and Federico 2013b; Goto et al. 2013). Nevertheless, POM remains

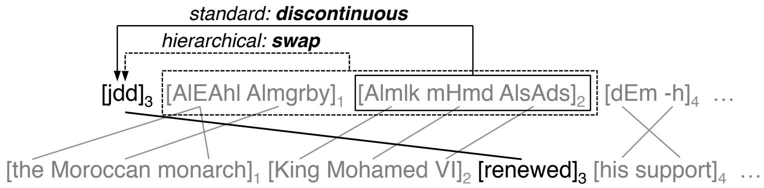

standard: discontinuous
hierarchical: swap
[jdd]3 [AlEAhl Almgrby]1 [Almlk mHmd AlsAds]2 [dEm -h]4 …
[the Moroccan monarch]1 [King Mohamed VI]2 [renewed]3 [his support]4 …

Figure 3
Phrase orientation example for the phrase pair *[jdd]-[renewed]*: the standard model detects a discontinuous orientation with respect to the last translated phrase (2) whereas the hierarchical model detects a swap with respect to the block of phrases (1-2).

the most widely used type of phrase-based reordering model and is considered a necessary component of PSMT baselines in any language pair. In particular, two variants of POM deserve further attention because of their notable effect on translation quality: hierarchical POM and sparse phrase orientation features.

Hierarchical phrase orientation models, or simply **hierarchical reordering models** (HRM) (Galley and Manning 2008), improve the way in which the orientation of a new phrase pair is determined: Already translated adjacent blocks are merged together to form longer phrases around the current one. For instance in Figure 3, HRM merges phrases 1 and 2 into a large phrase pair *[AlEahl ... AlsAds]-[The ... VI]* and consequently assigns a swap, instead of discontinuous orientation, to *[jdd]-[renewed]*. As a result, orientation assignments become more consistent across hypotheses with different phrase segmentations.

Rather than training a reordering model by relative frequency or maximum entropy and using its score as one dense feature function, Cherry (2013) introduces **sparse phrase orientation features** that are directly added to the model score during decoding (cf. Equation (1)) and optimized jointly with all other SMT feature weights. Effective sparse reordering features can be obtained by simply coupling a phrase pair's orientation with the first or last word (or word class) of its source and target side (Cherry 2013), or even with the whole phrase pair identity (Auli, Galley, and Gao 2014).

2.2 *n*-gram Based SMT

n-gram based SMT (Casacuberta and Vidal 2004; Mariño et al. 2006) is a string-based alternative to PSMT. In this framework, smoothed *n*-gram models are learned over sequences of minimal translation units (called **tuples**), which, like phrase pairs, are pairs of word sequences extracted from word-aligned parallel sentences. Tuples, however, are typically shorter than phrase pairs and are extracted from a unique, *monotonic* segmentation of the sentence pair. Thus, the problem of spurious phrase segmentation is avoided but non-local reordering becomes an issue. For instance, in Figure 2, a monotonic phrase segmentation could be achieved only by treating the large block *[jdd ... AlsAds]-[The ... renewed]* as a single tuple. Reordering is then addressed by "tuple unfolding" (Crego, Mariño, and de Gispert 2005): that is, during training the source words of each translation unit are rearranged in a target-like order so that more, shorter tuples can be extracted. At test time, input sentences have to be *pre-ordered* for translation. To this end, Crego and Mariño (2006) propose to precompute a number of likely permutations of the input using POS-based rewrite rules learned during tuple unfolding. The reorderings thus obtained are used to extend the search graph of a monotonic decoder.[8] Reordering is often considered as a shortcoming of *n*-gram–based SMT as reordering decisions are largely decoupled from decoding and mostly based on source-side information.

2.3 Tree-Based SMT

The SMT frameworks discussed so far learn direct correspondences between source and target words or phrases, treating reordering as a sequential process. This flat representation is fairly successful for some language pairs, although in others, reordering is more naturally described as a hierarchical process where small, locally reordered blocks become the elements of recursively larger reordered blocks. Concretely, in our running example (Figure 2), a hierarchical or tree-based approach would make it possible to first translate and reorder small blocks such as *[AlEahl Almgrby]* and *[Almlk mHmd AlsAds]*, then merge them to compose a larger block that gets reordered *as a whole* with respect to the verb *jdd*, and so forth. The degree of generalization at each level would then depend on how blocks are represented (e.g., by their lexical content, by a tag denoting the block's syntactic category, or by a generic symbol).

Tree-based approaches are largely inspired by syntactic parsing, but not all in the same way: Some model translation as the transformation of trees produced by monolingual parsers trained on syntactic treebanks (Section 2.3.1), whereas others extract a bilingual translation grammar directly from word-aligned parallel text without using any syntactic information (Section 2.3.2). Non-syntactic bilingual translation grammars may still be enriched with syntactic information—for instance, in the form of soft constraints (Section 2.3.3).

All tree-based frameworks crucially differ from PSMT and other string-based frameworks with respect to reordering: Whereas PSMT considers all input permutations that do not violate general reordering constraints and then scores them with separate reordering models, tree-based systems model reordering jointly with translation and, during decoding, only (or mostly) explore input permutations that are licensed by the learned translation model.

Most modern tree-based approaches fall under the general formulation of SMT, which scores translation hypotheses by a linear combination of feature functions (see Equation (1)), with a translation model (or grammar) and a target language model as core features. Tree-based decoding is usually performed by a chart-parsing algorithm with beam search and integrated target language model. Hence, the target sentence is not produced from left to right as in string-based SMT, but bottom–up according to a tree derivation order.

2.3.1 Syntax-Based SMT. An important motivation for using syntax in SMT is that reordering among natural languages very often involves the permutation of whole syntactic constituents (e.g., Fox 2002). For instance, in our running example (Figure 2), knowing the span of the Arabic subject would be enough to predict the reordering of the verb for translation into English.

Syntax-based SMT encompasses a variety of frameworks that use syntactic annotation either on the source or on the target language, or both. So-called **tree-to-string** methods (Huang, Knight, and Joshi 2006; Liu, Liu, and Lin 2006) use a given input sentence parse tree to restrict the application of translation/reordering rules to word

[8] More pre-ordering techniques will be discussed in Section 2.4.

spans that coincide with syntactic constituents of specific categories. For instance, the swap of *Alr}ys Alfrnsy* may only be dictated by a rule applying to noun phrases composed of a noun and an adjective. On the other hand, **string-to-tree** methods (Yamada and Knight 2002; Galley et al. 2004; Marcu et al. 2006; Shen, Xu, and Weischedel 2010) use syntax as a way to restrict translation hypotheses to well-formed target language sentences—ruling out, for instance, a translation that fails to reorder the translated verb *renewed* with respect to its subject. Using syntax on both source and target sides (**tree-to-tree**) (Imamura, Okuma, and Sumita 2005; Ding and Palmer 2005; Smith and Eisner 2006; Watanabe, Tsukada, and Isozaki 2006; Zhang et al. 2008) has proven rather difficult in practice due to the complexity of aligning potentially very different tree topologies and to the large size of the resulting translation grammars. Moreover, the need for high-quality parsers in both language sides seriously limits the applicability of this approach.

Syntax-based SMT approaches also differ in the formalism they use to represent the trees. Those based on phrase structure (constituency) grammars typically comply with the principle that each translation/reordering rule should match a complete constituent, whereas those based on dependency grammars opt for a more flexible use of structure. For example, in **string-to-dependency SMT** (Shen, Xu, and Weischedel 2010) rules can correspond to partial constituents but must be either a single rooted tree, with each child being a complete sub-tree, or a sequence of siblings, each being a complete sub-tree. Partial dependency rules are then combined during decoding, which means that not all reordering decisions are governed by the translation model.

An even more flexible use of structure is advocated by the **treelet-based SMT** framework (Quirk, Menezes, and Cherry 2005), where translation rules can correspond to any connected subgraph of the dependency tree (i.e., treelet). As illustrated by Figure 4, treelet pairs are extracted from pairs of source dependency parse tree and target-side projected trees. Treelets can be seen as phrases that are not limited to sets of adjacent words, but rather to sets of words that are connected by dependency relations, which in turn make it possible to learn non-local reordering patterns. As reordering

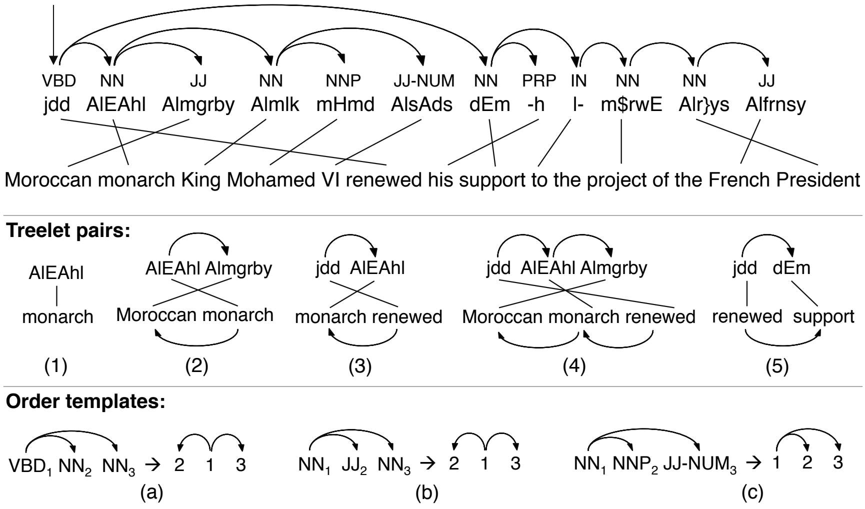


Figure 4
Examples of treelet pairs and order templates extracted from a word-aligned sentence pair and its source-side dependency parse tree. The projected tree for the whole target sentence is not shown due to space limitations.

decisions are only partially governed by the translation model, treelet-based SMT benefits from additional model components specifically dedicated to reordering. For example, in Figure 4, treelet pair (3) determines the swapping of *jdd* and *AlEAhl* but does not specify the ordering of *dEm*, which is also a child of *jdd*. Hence, during decoding, all possible reorderings of the unmatched children are considered and scored by a separate discriminative model, predicting the position of a child node (or modifier *m*) relative to its head *h*, given lexical, POS, and positional features of *m* and *h*. Reordering modeling is thus largely decoupled from lexical selection, which makes the model very flexible but results in a very large search space and high risk of search errors. To address this issue, Menezes and Quirk (2007) introduce another mechanism to complement treelet reordering: namely, dependency order templates. An order template is an unlexicalized rule specifying the reordering of a node and all its children based on their POS tags. For instance, in Figure 4, treelet pair (3) may be combined with template (a) to specify the order of the child *dEm*. For each new test sentence, matching treelet pairs and order templates are combined to construct lexicalized translation rules for that sentence and, finally, decoding is performed with a chart parsing algorithm.

We will now discuss SMT frameworks that model translation as a process of parallel parsing of the source and target language via a synchronous grammar.

2.3.2 Tree-Based SMT Without Syntax. The idea of extracting bilingual translation (i.e., synchronous) grammars directly from word-aligned parallel data originates in early work on ITG by Wu (1996, 1997).

In a more mature approach, **hierarchical phrase-based SMT (HSMT)** (Chiang 2005), the translation model is a probabilistic synchronous context-free grammar (SCFG) whose rules can correspond to arbitrary (i.e., nonsyntactically motivated) phrases labeled by only two generic non-terminal symbols (X or S). As shown in Figure 5, HSMT translation rules can either include a mix of terminals and non-terminals capturing reordering patterns and discontinuities (rules 1–4), or only terminals (rules 7–10) basically corresponding to phrase pairs in string-based PSMT. Finally, the so-called glue rules (5–6) are always added to the grammar to combine translated blocks in a monotone fashion regardless of their content. As in PSMT, extracted translation rules may not exceed a certain length and rule scores are obtained using maximum likelihood

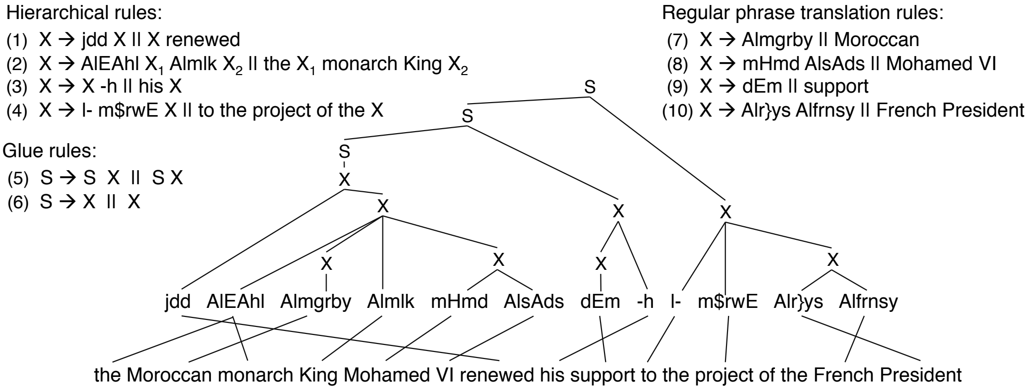

Hierarchical rules:
(1) X → jdd X || X renewed
(2) X → AlEAhl X_1 Almlk X_2 || the X_1 monarch King X_2
(3) X → X -h || his X
(4) X → l- m$rwE X || to the project of the X
Glue rules:
(5) S → S X || S X
(6) S → X || X
Regular phrase translation rules:
(7) X → Almgrby || Moroccan
(8) X → mHmd AlsAds || Mohamed VI
(9) X → dEm || support
(10) X → Alr}ys Alfrnsy || French President
S
S
S
X
X
X
X
X
X
X
X
X
jdd AlEAhl Almgrby Almlk mHmd AlsAds dEm -h l- m$rwE Alr}ys Alfrnsy
the Moroccan monarch King Mohamed VI renewed his support to the project of the French President

Figure 5
Possible derivation of a word-aligned sentence pair and corresponding hierarchical phrase-based translation grammar. The target-side tree is not represented due to space limitations.

estimation. Crucially, swapping adjacent phrases with no lexical evidence ($\mathrm{X} \rightarrow \mathrm{X}_1\mathrm{X}_2||\mathrm{X}_2\mathrm{X}_1$) is not allowed by standard HSMT grammars; therefore reordering can only be triggered by at least partially lexicalized translation rules. This is a major difference with respect to most syntax-based approaches, where reordering can be captured by rules containing only labeled non-terminals (e.g., $\mathrm{S} \rightarrow \mathrm{NP\ VP}\ ||\ \mathrm{VP\ NP}$). This means that, for instance, the reordering pattern learned by our example HSMT grammar (Figure 5, rule 1) may only be used to reorder the specific verb form *jdd (renewed)* in subsequent test sentences. Thus, HSMT is likely to work better for languages where the syntactic role of phrases is mostly expressed by separate function words (e.g., Chinese) than for languages where this information is largely conveyed by word inflection (e.g., Russian).

Although hierarchical models are inherently capable of dealing with complex and recursive reordering patterns, in practice many translation rules are noisy or based on limited context. To limit search complexity, a constraint is imposed on the maximum number of source words that may be covered by a non-terminal symbol during decoding (**span constraint**). This parameter is typically set to 10 or 15 words, as wider spans result in prohibitively slow decoding and lower translation quality. For these reasons, a number of extensions to the original HSMT framework have been proposed with the specific goal of better handling complex reordering phenomena.

Shallow-*n* grammars (de Gispert et al. 2010) can be used to refine the reordering space of HSMT according to the reordering characteristics of a specific language pair. For instance, as shown in Figure 6, an Arabic–English HSMT grammar is extended with an additional non-terminal symbol X0 that can only generate fully lexicalized phrases, thereby disallowing recursive nesting of hierarchical rules (shallow-1 grammar). To account for the movement of large word blocks, other new non-terminals M^k allow for the monotonic generation of *k* non-terminals X0. While defining a much smaller search space than the original HSMT grammar, the resulting shallow grammar can capture the long-range reordering of our running example even in the likely absence of a rule covering the whole subject span (i.e., in rule 2 in Figure 5).

In related work specifically addressing the issue of long-range reordering, Braune, Gojun, and Fraser (2012) propose relaxing the span constraint only for specific types of hierarchical rules that are more likely to capture long, reordering patterns in German–English. For instance, rules whose source side starts with at least one terminal followed by one non-terminal and ends with at least one terminal (t^+ X t^+) can capture the pattern

Regular phrase translation rules:
(1) X0 → AlEAhl Almgrby || the Moroccan monarch
(2) X0 → Almlk || King
(3) X0 → mHmd AlsAds || Mohamed VI
(4) X0 → dEm || support
Monotonic block rules:
(5) $\mathrm{M}^2 \rightarrow \mathrm{X0}_1\ \mathrm{X0}_2\ ||\ \mathrm{X0}_1\ \mathrm{X0}_2$
(6) $\mathrm{M}^3 \rightarrow \mathrm{X0}\ \mathrm{M}^2\ ||\ \mathrm{X0}\ \mathrm{M}^2$
Hierarchical rules:
(7) $\mathrm{X1} \rightarrow$ jdd M^3 || M^3 renewed
(8) X1 → X0 -h || his X0
Glue rules:
(9) S → S X1 || S X1
(10) S → X1 || X1

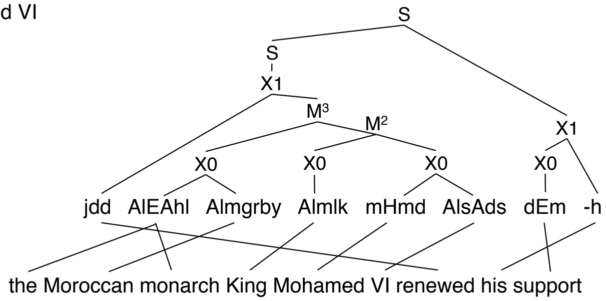

S
S
X1
M3
M2
X1
X0
X0
X0
X0
jdd AlEAhl Almgrby Almlk mHmd AlsAds dEm -h
the Moroccan monarch King Mohamed VI renewed his support

Figure 6
Example of shallow-1 HSMT grammar with monotonic non-terminals M^k. The target-side tree is not represented due to space limitations.

'finite-auxiliary-verb X participle' (e.g., *ist X gestiegen/has increased X*) with very wide X spans.

Mylonakis and Sima'an (2010) separate the modeling of local reordering (captured by fully lexicalized phrase-pair emission rules) from the modeling of higher-order recursive reordering (captured by ITG-style non-lexicalized binary rules). Instead of a single non-terminal X, three different **reordering-based labels** are used, according to the reordering pattern in which they participate: X for monotonic rules; XSL and XSR for the first and second symbol, respectively, of swapping rules. Thus reordering decisions are conditioned on the phrase pair's content, rather than its lexical context as in HSMT. More fine-grained non-terminals are introduced by Maillette de Buy Wenniger and Sima'an (2014) to also capture the relation of a phrase pair's reordering with respect to the parent phrase that contains it.

Rather than relabeling non-terminals, other work incorporates reordering-specific models as additional feature functions. He, Meng, and Yu (2010) add to their HSMT grammar the generic phrase swapping rule ($\mathrm{X} \rightarrow \mathrm{X}_1\mathrm{X}_2||\mathrm{X}_2\mathrm{X}_1$) and use a maximum-entropy classifier to predict whether two neighboring phrases should be swapped or not during decoding. Rather than conditioning the decision on the whole phrase pair, the classifier uses features extracted from it, such as first and last word (or POS tag) of the source and target side. A similar model was first developed by Xiong, Liu, and Lin (2006) for simpler phrase translation models (i.e., without discontinuities) based on ITG. Li, Liu, and Sun (2013) use recursive autoencoders (Socher et al. 2011) to assign vector representations to the neighboring phrases given as input to the ITG classifier, thereby avoiding manual feature engineering but affecting hypothesis recombination and decoding speed. Nguyen and Vogel (2013) and Huck et al. (2013) successfully integrate the distortion cost feature function and phrase orientation models initially designed for string-based PSMT into a chart-based HSMT decoder.

Finally, Setiawan, Kan, and Li (2007) observe that, in languages like Chinese and English, function words provide important clues on the grammatical relationships among phrases. Consequently, they introduce a SCFG where *function words* (approximated by high-frequency words) are the only lexicalized non-terminals guiding phrase reordering. Based on the same intuition, Setiawan et al. (2009) augment a HSMT system with a function-word ordering model that predicts, for any pair of translation rules, which one should dominate the other in the hierarchical structure, based on the function words that they contain.[9]

2.3.3 Tree-Based SMT with Soft Syntactic Constraints. We have discussed SMT frameworks where the translation model is fully based on the syntactic parse tree of the source or target sentence (Section 2.3.1) or where syntax is not used at all (Section 2.3.2). A third line of work bridges between these two by exploiting syntactic information in the form of soft constraints while operating with a synchronous translation grammar extracted from non-parsed parallel data.

Chiang (2005) first experimented with a feature function rewarding translation rules applied to full syntactic constituents (**constituent feature**). Although this initial attempt did not appear to improve translation quality, Marton and Resnik (2008) further elaborated the idea and proposed a series of finer-grained features distinguishing among

[9] Two other models utilizing function words as the *anchors* of global reordering decisions are proposed in Setiawan et al. (2013) and Setiawan, Zhou, and Xiang (2013). Although integrated in a syntax-based system (Shen, Xu, and Weischedel 2010), these models are in principle applicable to other SMT frameworks such as HSMT.

constituent types (VP, NP, etc.), eventually leading to better performance. Gao, Koehn, and Birch (2011) extract two reordering-related feature functions from source dependency parse trees: (i) The **dependency orientation model** predicts whether the relative order of a source word and its head should be reversed during translation. This is trained as a maximum-entropy classifier using the words and their dependency relation type as features. (ii) The **dependency cohesion penalty** fires whenever a word and its head are translated separately (i. e., by different translation rules), thereby measuring derivation well-formedness. Because long-range reordering tends to happen closer to the root and local reordering closer to the leaves, a distinction is made between words occurring at different depths of the dependency tree leading to a number of subfeatures. In this way, the tuning process can decide how important or reliable feature scores coming from different levels of the parse tree are. Huang, Devlin, and Zbib (2013) worked instead with constituency parses and trained a classifier to predict whether the order of any two sibling constituents in the input tree should be reversed or maintained during translation. The classifier is trained by maximum entropy, using a number of syntactic features and used during decoding at the word level: that is, each pair of input words inherit the orientation probabilities of the constituents that cover them, respectively.

Syntactic annotation has also been used to refine non-terminal SCFG labels, potentially leading to better reordering choices. In Zollmann and Venugopal (2006) and Mylonakis and Sima'an (2011), labels indicate whether a phrase corresponds to a syntactic constituent or to part of it, as well as the constituent type, relatively to a target or source parse tree, respectively. Moreover, Mylonakis and Sima'an treat the phrase-pair category as a latent variable and let their system learn reordering distributions over multiple labels per span (generic X or source-syntax based like NP, VBZ+DT, etc.). Li et al. (2012) use source dependency annotation to refine non-terminal symbols with syntactic head information. More specifically, given a hierarchical phrase, its type is obtained by concatenating the POS tags of the exposed heads it contains on the source side, where an exposed head is a word dominated by a word outside the phrase. Like He, Meng, and Yu (2010), Li et al. (2012) also allow adjacent phrases to swap, but instead of introducing a separate orientation model, they rely on rule translation probabilities based on the refined non-terminals to guide reordering.

2.4 Word Reordering as Pre- (or Post-) Processing

Given the complexity of solving word reordering during the decoding process, a productive line of research has focused on decoupling reordering decisions from translation decisions. These approaches aim at arranging words in a target-like order either on the input, *before* translating, or on the output, *after* translating. Thus, word reordering is solved as pre- or post-processing (i.e., **pre-ordering** or **post-ordering**) in a monolingual fashion and with unconstrained access to the whole sentence context. Figure 7 (Sudoh et al. 2011) illustrates the workflows of pre- and post-ordering approaches as opposed to standard SMT.

2.4.1 Main Pre-ordering Strategies. A large number of pre-ordering strategies have been proposed. As a first classification, we divide them into deterministic, non-deterministic, and hybrid. **Deterministic pre-ordering** aims at finding a single optimal permutation of the input sentence, which is then translated monotonically or with a low distortion limit (Nießen and Ney 2001; Xia and McCord 2004; Collins, Koehn, and Kucerova 2005; Popović and Ney 2006; Costa-jussà and Fonollosa 2006; Wang, Collins, and Koehn 2007;

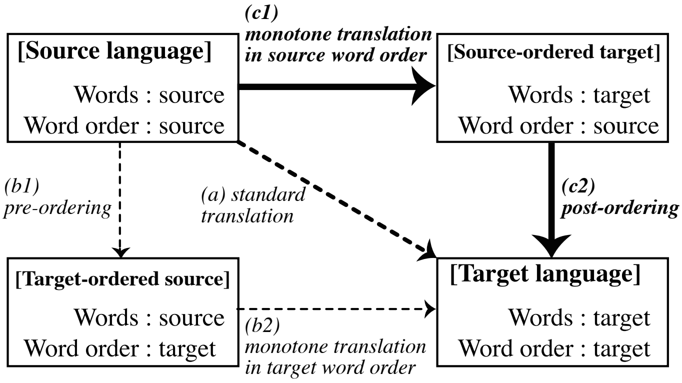


Figure 7
Typical workflows of standard, pre-ordering, and post-ordering approaches to SMT. Taken from Sudoh et al. (2011).

Habash 2007; Li et al. 2007; Tromble and Eisner 2009; Xu et al. 2009; Genzel 2010; Isozaki et al. 2010b; Yeniterzi and Oflazer 2010; Khalilov and Fonollosa 2011; Khalilov and Sima'an 2011; Visweswariah et al. 2011; Gojun and Fraser 2012; Yang et al. 2012; Lerner and Petrov 2013; Jehl et al. 2014).[10] **Non-deterministic pre-ordering** encodes multiple alternative reorderings into a word lattice and lets a monotonic (usually n-gram–based) decoder choose the best path according to its models (Zens, Och, and Ney 2002; Kanthak et al. 2005; Crego and Mariño 2006; Zhang, Zens, and Ney 2007; Rottmann and Vogel 2007; Crego and Habash 2008; Elming and Habash 2009; Niehues and Kolss 2009). A **hybrid approach** is adopted by Bisazza and Federico (2010) and Andreas, Habash, and Rambow (2011): Rules are used to generate multiple likely pre-orderings, but only for specific language phenomena that are responsible for difficult (long-range) reordering patterns. The sparse reordering lattices produced by these techniques are then translated by a decoder performing additional phrase-based reordering. In a follow-up work, Bisazza and Federico (2012) introduce another way to encode multiple pre-orderings of the input: Instead of generating a word lattice, pre-computed permutations are represented by a **modified distortion matrix** so that lower distortion costs or "shortcuts" are permitted between selected pairs of input positions.

Pre-ordering methods can also be classified by the kind of pre-ordering rules that they apply: that is, manually written based on linguistic knowledge, or automatically learned from data. We now discuss each of them in detail.

2.4.2 Linguistic Knowledge–Based Pre-ordering. In these approaches, manually written rules determine the transformation of input syntax trees (Collins, Koehn, and Kucerova 2005; Wang, Collins, and Koehn 2007; Xu et al. 2009; Isozaki et al. 2010b; Yeniterzi and Oflazer 2010; Gojun and Fraser 2012; Andreas, Habash, and Rambow 2011) or the permutation of shallow syntactic chunks in a sentence (Hardmeier, Bisazza, and Federico 2010; Durgar El-Kahlout and Oflazer 2010; Bisazza, Pighin, and Federico 2012). In an early example of syntax-based pre-ordering, Collins, Koehn, and Kucerova (2005) propose a set of six rules aimed at arranging German sentences in English-like order. The rules address the position of verbs, verb particles, and negation particles, and they

[10] Li et al. (2007) experiment with a small number of n-best pre-orderings given as alternative inputs to the SMT system.

are applied to constituency parse trees. Following a similar approach, Gojun and Fraser (2012) develop a set of rules for the opposite translation direction (English-to-German). Xu et al. (2009) instead propose a simple set of dependency-based rules to pre-order English for translation into subject-object-verb (SOV) languages, which is shown to be effective for Korean, Japanese, Hindi, Urdu, and Turkish. Isozaki et al. (2010b) obtain even better results in an English-to-Japanese task using only one pre-ordering rule (i.e., head finalization) with a parser annotating syntactic heads.

2.4.3 Data-Driven Pre-ordering. This kind of model is learned from sets of pairs $(\mathbf{f}, \mathbf{f}')$ where $\mathbf{f}$ is a source sentence and $\mathbf{f}'$ is its reference permutation (pre-ordering) inferred from a reference translation $\mathbf{e}$ via a word-level alignment.[11] These approaches typically require some form of linguistic annotation of the source language, such as syntactic parse trees (Xia and McCord 2004; Habash 2007; Li et al. 2007; Elming and Habash 2009; Genzel 2010; Khalilov and Fonollosa 2011; Khalilov and Sima'an 2011; Yang et al. 2012; Lerner and Petrov 2013; Jehl et al. 2014), shallow syntax chunks (Zhang, Zens, and Ney 2007; Crego and Habash 2008), or POS labels (Crego and Mariño 2006; Rottmann and Vogel 2007; Niehues and Kolss 2009; Tromble and Eisner 2009; Visweswariah et al. 2011).

Among the first examples of data-driven tree-based pre-ordering, Xia and McCord (2004) propose a method to automatically learn reordering patterns from a dependency-parsed French–English bitext, using a number of heuristics. While source-side parses are required by their method, target-side parses are optionally used to provide additional constraints during rule extraction. Habash (2007) extracts pre-ordering rules from an Arabic–English parallel corpus dependency-parsed on the source side. In both these works, pre-ordering rules are applied in a deterministic way to preprocess both training and test data. Following a discriminative modeling approach, Li et al. (2007) train a maximum-entropy classifier to pre-order each node with at most three children in the source constituency parse, using a rich set of lexical and syntactic features. Lerner and Petrov (2013) extend this work to pre-order nodes with more children (up to seven on either side of the head) using a cascade of classifiers: first, decide the order of each child relative to the head, then decide the order of left children and that of the right children. As training separate classifiers for each number of children is prone to sparsity issues, Jehl et al. (2014) build a single logistic regression model to predict whether any two sibling nodes should be swapped or not. Then, for each node in the tree, they search for the best permutation of all its children given the pairwise scores produced by the model, using a depth-first procedure. Yang et al. (2012) treat the permutation of each node's children as a ranking problem and model it with ranking support vector machines. As an alternative to deterministic pre-ordering, they also propose using the predicted source permutation to generate soft constraints for the SMT decoder: that is, a penalty that fires whenever the decoder violates the predicted pre-ordering. A tighter integration between source pre-ordering and source-to-target translation is proposed by Dyer and Resnik (2010). In their approach, optimal source pre-orderings $(\mathbf{f}')$ are treated as a latent variable in an end-to-end translation model and the parameters of the tree permutation model are learned directly from parallel data. At test time, alternative permutations of the input tree are encoded as a **source reordering forest**, which is then translated by a finite-state phrase-based translation model.

[11] Various heuristics have been proposed to convert a word alignment set into a sentence permutation (Birch, Osborne, and Blunsom 2010; Feng, Mauser, and Ney 2010; Visweswariah et al. 2011).

Examples of pre-ordering based on shallow syntax include Zhang, Zens, and Ney (2007) and Crego and Habash (2008). In these approaches, automatically extracted chunk pre-ordering rules are used to generate a word reordering lattice of the input sentence, which is then translated by a monotonic phrase or *n*-gram–based decoder.

In Costa-jussà and Fonollosa (2006), pre-ordering is learned by training a monolingual *n*-gram based SMT system at the level of word clusters. In Tromble and Eisner (2009), pre-ordering is cast as a permutation problem and solved by a model that estimates the probability of reversing the relative order of any two input words based on their distance as well as lexicalized and POS-based features. In a related work, Visweswariah et al. (2011) obtain smaller models and better results by learning the cost of a given input word appearing right after another, as opposed to anywhere after it (cf. source word-after-word reordering models described in Section 2.1).

2.4.4 On the Limitations of Syntax-based Pre-ordering. Syntax is often regarded as the most effective way to inform reordering in translation. However, empirical work has shown that the success of syntax-based pre-ordering methods can be severely limited by (i) the reachability of reference permutations when parse trees are used to constrain the pre-ordering model, and (ii) the quality of the parser used to learn and apply a pre-ordering model.

With regard to the constraints imposed by syntactic trees (i), Khalilov and Sima’an (2012) conducted oracle pre-ordering experiments across various language pairs. Their results consistently showed that final translation quality was highest by far when no syntactic constraint was imposed on pre-ordering (oracle string). On the contrary, only allowing permutations of siblings of the source parse tree (oracle tree) gave the smallest improvement. Only some of this loss could be recovered by applying specific modifications to the tree before extracting the optimal permutation (oracle modified tree).

With regard to parser accuracy (ii), Green, Sathi, and Manning (2009) analyzed two state-of-the-art parsers (Bikel 2004; Klein and Manning 2003) and reported F-measures of only 55% to 56% at the sub-task of detecting Arabic NP subjects in verb-initial clauses. Similar results were observed by Carpuat, Marton, and Habash (2010) using a dependency parser (Nivre, Hall, and Nilsson 2006). The same study also showed that the correct pre-ordering for Arabic–English translation could not be safely predicted even from gold standard parses, partly because of syntactic transformations occurring during translation. From a manual analysis of their English–German system, Gojun and Fraser (2012) reported that about 10% of the English clauses were wrongly pre-ordered, mostly from source sentence parsing errors. Howlett and Dras (2011) analyzed a reimplementation of the German pre-ordering method of Collins, Koehn, and Kucerova (2005) and found that results could be affected—or even cancelled out—by many factors including choice of training data, quality of the parser, as well as order of the target language model and type of reordering model used during decoding.

Rather than relying on supervised parsers trained on golden treebanks, specific parsers can be induced directly from non-annotated parallel text. In DeNero and Uszkoreit (2011), source sentence reorderings are first inferred from the word alignment with the target translation. Then, a binary parsing model is trained to maximize the likelihood of source trees that can generate such reorderings. Finally, a pre-ordering model is trained to permute each node in the tree. Evaluated on the English–Japanese language pair, this method almost equals the performance of a pre-ordering method based on a supervised parser. Neubig, Watanabe, and Mori (2012) follow a similar approach but build a single ITG-style pre-ordering model treating the parse tree as a latent variable. In the target self-training method of Katz-Brown et al. (2011), a baseline treebank-trained parser is used to produce *n*-best parses of a parallel corpus's source side. Then, the parses resulting in the most accurate pre-ordering after application of a dependency-based pre-ordering rule set (Xu et al. 2009) are added to the treebank data and used to re-train the baseline parser.

2.4.5 Post-ordering. A somewhat smaller line of research has instead treated reordering as post-processing. In Bangalore and Riccardi (2000) and Sudoh et al. (2011), target words are reordered after a monotonic translation process. Other work has focused on rescoring a set of *n*-best translation candidates produced by a regular PSMT decoder—for instance, by means of POS-based reordering templates (Chen, Cettolo, and Federico 2006) or word-class specific distortion models (Gupta, Cettolo, and Federico 2007). Chang and Toutanova (2007) use a dependency tree reordering model to generate *n* alternative orders for each 1-best sentence produced by the SMT system. Each set of *n* sentence reorderings is then reranked using a discriminative model trained on word bigram features and standard word reordering features (i.e., distance or orientation between consecutively translated input words).

Focusing on Japanese-to-English translation, Sudoh et al. (2011, 2013) proposed to "translate" foreign-order English into correct-order English using a monolingual phrase-based (Sudoh et al. 2011) or syntax-based (Sudoh et al. 2013) SMT system trained for this specific subtask.[12] The underlying motivation is that, while English-to-Japanese is well handled by pre-ordering with the aforementioned head-finalization rule (Isozaki et al. 2010b), it is much harder to predict the English-like order of Japanese constituents for Japanese-to-English translation. Post-ordering addresses this issue by generating head-final English (HFE) sentences that are used to create a HFE-to-English parallel corpus. Goto, Utiyama, and Sumita (2012, 2013) solve post-ordering by parsing the HFE sentences into binary trees annotated with both syntactic labels and ITG-style monotone/swap labels. Hayashi et al. (2013) improve upon this work with a shift-reduce parser that efficiently integrates non-local features like *n*-grams of the post-ordered string.

Also related to post-ordering is the work on right-to-left or **reverse decoding** by Watanabe and Sumita (2002), Finch and Sumita (2009), and Freitag et al. (2013). Here, the target sentence is built up from the last word to the first, thereby altering language model context and reordering search space. Finch and Sumita obtain best results on a wide range of language pairs by combining the outputs of standard and reverse decoding systems.

3. Evaluating Word Reordering in Statistical Machine Translation

Because there are innumerable ways to correctly render a source sentence's meaning in the target language, automatically evaluating translation quality is a complex problem. Generally, SMT systems are judged by the extent to which their outputs resemble a set of reference translations produced by different human translators. Despite relying on a very rough approximation of language variability, this approach provides SMT researchers with fast automatic metrics that can guide, at least in part, their steps towards improvement. Besides, fast evaluation metrics are used to automatically tune SMT

[12] Note the similarity to the pre-ordering approach of Costa-jussà and Fonollosa (2006), except that here the monolingual SMT process is applied to the target language after a monotonic translation phase.

feature weights on a development corpus—for instance, by means of minimum error rate training procedures (Och 2003). The design of MT evaluation metrics correlating with human judgments is an active research area. Here we briefly survey two widely used general-purpose metrics, BLEU and METEOR, and then describe in more detail a number of reordering-specific metrics.

3.1 General-Purpose Metrics

BLEU (Papineni et al. 2001) is a lexical match–based score that represents the de facto standard for SMT evaluation. Here, proximity between candidate and reference translations is measured in terms of overlapping word n-grams, with n typically ranging from 1 to 4. For each order n a **modified precision** score (see Papineni et al. [2001] for details) is computed on the whole test set and combined in a geometric mean. The resulting score is then multiplied by a **brevity penalty** that accounts for length mismatches between reference and candidate translations. Al-Onaizan and Papineni (2006) use BLEU to measure word order similarity between two languages: that is, by computing the BLEU score between the original target sentence **e** and a source-like permutation of **e**. Using n-grams, though, is a limited solution to the problem of word ordering evaluation. First, because only exact surface matches are counted, without any consideration of morphology or synonymy. Second, because the absolute positioning of words in the sentence is not captured, but only their proximity within a small context.

The former issue is addressed to some extent by **METEOR** (Banerjee and Lavie 2005), which relies on language-specific stemmers and synonymy modules to go beyond the surface-level similarity. As for word order, METEOR treats it separately with a **fragmentation penalty** proportional to the smallest number of chunks that the hypothesis must be divided into to align with the reference translation. This quantity can be interpreted as the number of times that a human reader would have to "jump" between words to recover the correct translation order. However, no distinction is made between short and long-range reordering errors.

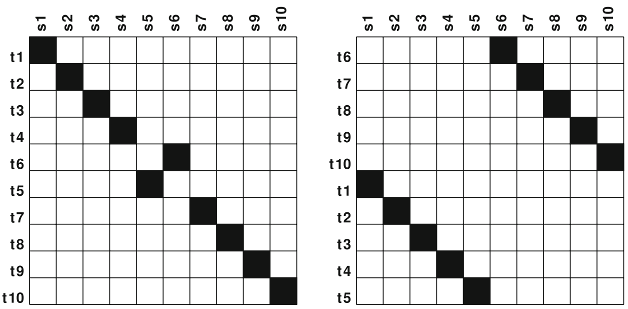
The weakness of BLEU and METEOR with respect to word order was demonstrated by Birch, Osborne, and Blunsom (2010) with a significant example that we report in Table 2. For simplicity, the example assumes that the reference order is monotonic and that hypotheses and reference translations contain exactly the same words. According to both metrics, hypothesis (a) is worse than (b), although in (a) only two adjacent words are swapped whereas in (b) the two halves of the sentence are swapped.

3.2 Reordering-Specific Metrics

To overcome the aformentioned limitations, Birch, Osborne, and Blunsom (2010) propose directly measuring the similarity between the reorderings needed to reach the reference translations from the source sentence and those applied by the decoder to produce the candidate translation. In practice, this is done by first converting word alignments to permutations using simple heuristics to handle null and multiple alignments, and then computing a permutation distance among the resulting permutations. Among various metrics proposed in the paper, the square root of the Kendall's Tau was shown to be reliable and highly correlated with human judgments.

The normalized Kendall's tau distance K is originally a measure of disagreement between rankings. Given a set of n elements and two permutations π and σ, the K distance corresponds to the number of discordant pairs (i.e., pairs of elements whose

Table 2
Two example alignments and their respective BLEU and METEOR scores, assuming that the reference alignment is monotonic. The permutation resulting from the hypothesis alignment is reported under each matrix, where bullet points represent jumps between non-sequential indices. Taken from Birch (2011).



(1 2 3 4 • 6 • 5 • 7 8 9 10)
Example **(a)**

(6 7 8 9 10 • 1 2 3 4 5)
Example **(b)**

Example:	**(a)**	**(b)**
BLEU	61.80	81.33
METEOR	86.91	92.63

relative order differs in the two permutations) normalized by the total number of ordered element pairs:

$$K(\pi,\sigma) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{d}(i,j)}{\frac{1}{2}n(n-1)} \quad \text{where} \quad \mathbf{d}(i,j) = \begin{cases} 1 \text{ if } \pi_i < \pi_j \text{ and } \sigma_i > \sigma_j \\ 0 \text{ otherwise} \end{cases}$$

Birch, Osborne, and Blunsom (2010) further suggest extract the square root of *K* to obtain a function that is more discriminative on lower distance ranges, (i.e., for translations that are closer to the reference word ordering). Finally, the **Kendall Reordering Score (KRS)**—a positive measure of quality ranging from 0 to 1—is computed by subtracting the latter quantity from one, and by multiplying the result by a brevity penalty (*BP*) that accounts for length mismatches between reference and candidate translations:

$$KRS(\pi,\sigma) = (1 - \sqrt{K(\pi,\sigma)}) \cdot BP$$

The *BP* definition corresponds to that of BLEU (Papineni et al. 2001) with the difference that, for KRS, it is computed at the sentence level. In case of multiple references, the one that yields the highest score for each test sentence is retained. Finally, the average of all sentence-level KRS scores gives the global KRS of the test set. The linear interpolation of KRS and BLEU (**LRscore**) can be successfully used to optimize the feature weights of a PSMT system, leading to translation outputs that are preferred by human annotators according to Birch and Osborne (2011).

In a related work, Bisazza and Federico (2013a) observe that some word classes, like verbs, are typically more important than others to determine the general structure of a sentence. Hence, they develop a word-weighted KRS variant that is more sensitive to the positioning of specific input words. Assuming that each input word f_i is assigned a weight λ_i, the original KRS formula is modified as follows:

$$\mathbf{d}_\lambda(i,j) = \begin{cases} \lambda_i + \lambda_j \text{ if } \pi_i < \pi_j \text{ and } \sigma_i > \sigma_j \\ 0 \quad\quad\; \text{otherwise} \end{cases}$$

For their evaluation of long reordering errors in Arabic–English and German–English, Bisazza and Federico (2013a) set the weights to 1 for verbs and 0 for all other words to only capture verb reordering errors. The resulting metric, **KRS-V**, rates a translation hypothesis as perfect when the translations of all source verbs are located in their correct position, regardless of the ordering of other words.

In a different approach called **RIBES**, Isozaki et al. (2010a) propose directly measuring the reordering occurring between the words of the hypothesis and those of the reference translation, thereby eliminating the need to word-align input and output sentence. A limitation of this approach is that only identical words contribute to the score. As a solution, the permutation distance is multiplied by a word precision score that penalizes hypotheses containing few reference words. Nevertheless, the resulting metric assigns different scores to hypotheses that differ in their lexical choice, but not in their word reordering.

Talbot et al. (2011) introduce yet another reordering-specific metric, called fuzzy reordering score (**FRS**) which, like the KRS, is independent from lexical choice and measures the similarity between a sentence's reference reordering and the reordering produced by an SMT system (or by a pre-ordering technique). However, whereas Birch, Osborne, and Blunsom (2010) used Kendall's tau between the two sentence permutations, Talbot et al. count the smallest number of chunks that the hypothesis permutation must be divided into to align with the reference permutation. This corresponds precisely to the fragmentation penalty of METEOR except that the alignment is performed between permutations and not between translations. Like METEOR, FRS makes no difference between short and long-range reordering errors (cf. Table 2).

Stanojević and Sima'an (2014b) argue for a hierarchical treatment of reordering evaluation, where word sequences can be grouped recursively into larger blocks. To this end, they factorize the output-reference reordering into a Permutation Tree (Zhang and Gildea 2007), whose nodes represent atomic permutations. Given this factorization, the counts of monotone (1 2) versus other permutation nodes—(2 1), (3 1 4 2), and so on—are used as features in a linear model of translation quality (**BEER**) trained to correlate with the human ranking of a set of MT system outputs. With reference to Table 2, the permutation trees of both hypotheses (a) and (b) would contain only one swapping node leading to the same reordering score. Stanojević and Sima'an (2014a) extend this work with a stand-alone reordering metric that considers all possible tree factorizations of a permutation (permutation forest) and that gives recursively less importance to lower nodes in the tree (i.e., covering smaller spans). Hierarchical permutation metrics are shown to better correlate with human judgments than string-based permutation metrics like Kendall's tau distance *K*.

4. Reordering Phenomena in Natural Languages

Understanding the complexity of reordering in a given language pair is key to selecting the right SMT models and to improving them. To date, word reordering phenomena in natural languages have mainly been analyzed from a quantitative perspective (Birch, Osborne, and Koehn 2008; Birch, Blunsom, and Osborne 2009). While measuring the *amount* of reordering is certainly important, understanding which *kinds* of reordering occur in a given language pair is also essential. To this end, we present a qualitative analysis of word reordering based on linguistic knowledge. More specifically, we draw on a large body of syntactic information collected by linguists from more than

1500 languages, and systematized in the World Atlas of Language Structures (WALS) (Dryer and Haspelmath 2011).[13]

Following the seminal work of language typologist Matthew S. Dryer, we describe the word order profile of a language by the canonical orders of its constituent sets (word order features). The resulting language pair classification is primarily based on the order of subject, object and verb, and further refined according to the order of several other element pairs, such as noun-adjective, verb-negation, and so forth. We then compare the word order features of several languages that were studied in the SMT field, and show that empirical results generally confirm the existing theoretical knowledge.

4.1 A Qualitative Analysis

The amount of word reordering found in a language pair is known to be a good predictor of SMT performance. Birch, Osborne, and Koehn (2008) considered three variables—reordering quantity, morphological complexity, and historical relatedness—and found the first to have the highest correlation with the BLEU scores of a standard PSMT system on a sample of 110 European language pairs. Birch, Blunsom, and Osborne (2009) further analyzed the distribution of different reordering widths in Arabic–English and Chinese–English, and the ability of two SMT approaches to model them. They found that the PSMT approach is more suitable for language pairs where most reordering is local (Arabic–English), while the hierarchical approach is stronger when medium-range reorderings are dominant (Chinese–English). Still, both PSMT and HSMT failed to capture most of the long-range reorderings found in the reference corpora.

These findings are indeed relevant to our work, but we believe there is also much to learn from theoretical linguistic knowledge. Moreover, a quantitative analysis can suffer from noise in the data, typically originating from automatic word alignments. Birch, Blunsom, and Osborne (2009) used manual word alignment in their study, but this kind of resource is available only for very few language pairs. Noise can also be due to what we can call *optional* reordering: Human translators often choose to restructure the sentence according to genre conventions or to their personal style, even when this is not required by the target language grammar. Here is an example:

Arabic sentence:

و طمأن بوش (55 سنة) الصحافيّن قبيل مغادرته البيت الابيض الى انه يشعر بانه في
حال " رائعة " و صحة " جيدة جدا " .

Literal translation:

Bush, aged 55, assured journalists before leaving the White House that he felt "great" and that his health was "very good".

Human translation:

Before leaving the White House, Bush, aged 55, assured journalists that he felt "great" and that his health was "very good".

As also noted by Fox (2002), this kind of reordering is not strictly necessary to produce accurate and fluent translations, but its occurrence in parallel corpora affects the automatic reordering measures.

[13] `http://wals.info`.

On the contrary, a qualitative analysis can profit from the extensive work done by linguists and grammaticians to abstract the fundamental properties of a language. In this section we draw largely on Dryer (2007) and on the sections of WALS devoted to word order (Dryer 2011, ch. 81–97, 143–144).

4.2 Word Order Profiles

The word order profile of a language is determined by the canonical order of its constituent sets, or word order features. In general, the basic or canonical order of a constituent set can be established by criteria of frequency (the most common), distribution (the one with the least restricted usage), or pragmatics (the neutral one) (Dryer 2007). Although some languages are said to have free (or flexible) order, it is often possible to detect one that is dominant and neutral. Consider, for instance, English, a SVO language where other orders are used, but only to achieve specific emphasis or topicalization effects:

(1) a. I saw the cat.
b. The cat, I saw.

However, there exist cases where no particular order can be defined as dominant. An example of mix-ordered constituent set in English is the pair noun and genitive:

(2) a. the tail of the cat
b. the cat's tail

Based on Dryer (2007) and on the availability of data points in the WALS, we have established a set of 13 core features to determine the word order profile of a language. For the purpose of describing word order differences between language pairs, we have divided the features into two broad categories: clause-level and phrase-level.[14] An English example for each feature is provided in Table 3.

4.2.1 Clause-Level Order Features.

- **Subject, Object, Verb** [WALS feature 81A]
The first and most important feature is the "ordering of subject, object, and verb in a transitive clause, more specifically declarative clauses in which both the subject and object involve a noun (and not just a pronoun)" (Dryer 2011). For instance, English and French are SVO languages, whereas Turkish is SOV. The distribution of main word order types in a large sample of world languages is given in Table 4. This feature is often used alone to denote the word order profile of a language, because it can be a good predictor of several other features.
- **Oblique or Adpositional Phrase** [WALS feature 84A]
This feature refers to the position of a phrase functioning as an adverbial modifier of the verb, relative to the position of the object and verb. For

[14] In this section, phrase is used in its traditional syntactic sense—i. e., a group of words forming a constituent—as opposed to the notion of data-driven phrase adopted by phrase-based SMT.

instance, English is VOX because it places oblique phrases after the verb and object.

- **Noun and Relative Clause** [WALS feature 90A]
  Order of the relative clause with respect to the noun it modifies.
- **Adverbial Subordinator and Subordinate Clause** [WALS feature 94A]
  Subordinators are used to link adverbial subordinate clauses to the main clause. They can take the form of verbal suffixes or separate words, such as the English subordinating conjunctions *when* and *because*.
- **Polar Question Particle** [WALS feature 92A]
  In many languages, polar (yes/no) questions are signaled by specific particles. This feature denotes their position in the sentence (not defined for English).
- **Content Question Phrase** [WALS feature 93A]
  Content questions are characterized by the presence of an interrogative word or phrase (e.g., *who*, *which one*). In some languages, like English, these are always placed at the beginning of the sentence. In some others, like Turkish, they take the position of the constituent they replace: For instance, the word '*ne*/what' replacing the object naturally occurs between subject and
  verb.
- **Negation and Verb** [WALS feature 143A]
  Order of the negative word or morpheme[15] with respect to the *main* verb. Note that more than one word or morpheme may be necessary to express negation (e.g., '*ne ... pas*' in French).

4.2.2 Phrase-Level Order Features.

- **Noun and Adpositions** [WALS feature 85A]
  Whether a language uses mainly prepositions or postpositions.
- **Noun and Genitive** [WALS feature 86A]
  Order of genitive or possessor noun phrase with respect to the head noun.
- **Noun and Adjective** [WALS feature 87A]
  Order of adjectives with respect to the noun they modify.
- **Noun and Demonstrative** [WALS feature 88A]
  Order of demonstrative words (e.g., *this*, *that*) or affixes with respect to the noun they modify.
- **Noun and Numeral** [WALS feature 89A]
  Order of cardinal number words with respect to the noun they modify.
- **Adjective and Degree Word** [WALS feature 91A]
  Order of degree words (e.g., *very*, *more*) with respect to the adjective they modify.

[15] Unlike the WALS, we do not distinguish between negative words and affixes for this feature.

Table 3
The word order profile of seven world languages. Language family and genus (Dryer 1989) are indicated in the header's first and second row, respectively. Sources: the World Atlas of Language Structures (Dryer and Haspelmath 2011). *Authors' knowledge. **Li (2008).

		Indo-European			*Afro-Asiatic*	*Altaic*	*Japanese*	*Sino-Tibetan*
		Germanic		*Romance*	*Semitic*	*Turkic*	*Japanese*	*Chinese*
	Features	**English**	**German**	**French**	**Arabic**	**Turkish**	**Japanese**	**Chinese**
Clause-level	Subject,Object,Verb [Tom] [chases] [Jerry]	S-V-O	S-V-O/ S-O-V	S-V-O	V-S-O/ S-V-O*	S-O-V	S-O-V	S-V-O
	Oblique Phrase [chases] [Jerry] [with a stick]	V-O-X	mixed	V-O-X	V-O-X	X-O-V	X-O-V	X-V-O
	Noun,RelClause [a stick] [that he stole]	N-Rel	N-Rel	N-Rel	N-Rel*	Rel-N	Rel-N	Rel-N
	Subordinator,Clause [because] [he was hungry]	Sub-C	Sub-C	Sub-C	Sub-C	C-Sub/ Sub-C	C-Sub	mixed**
	PolarQuest.Particle ∅ [did Tom steal it?]	*none*	*none*	initial	initial	final	final	final
	ContentQuest.Phrase [what] [did Tom steal?]	initial	initial	initial	initial*	other	other	other
	Negation,Verb he did [not] [steal]	Neg-V	Neg-V/ V-Neg	Neg-V-Neg/ V-Neg	Neg-V	V-Neg	V-Neg	Neg-V
Phrase-level	Noun,Adpositions [with] [a stick]	Adp-N	Adp-N	Adp-N	Adp-N	N-Adp	N-Adp	N-Adp/ Adp-N
	Noun,Genitive [Tom's] [stick]	N-Gen/ Gen-N	N-Gen	N-Gen	N-Gen	Gen-N	Gen-N	Gen-N
	Noun,Adjective [hungry] [Tom]	A-N	A-N	N-A	N-A	A-N	A-N	A-N
	Noun,Demonstrative [this] [stick]	Dem-N	Dem-N	Dem-N	Dem-N	Dem-N	Dem-N	Dem-N
	Noun,Numeral [two] [sticks]	Num-N	Num-N	Num-N	Num-N	Num-N	Num-N	Num-N
	Adjective,DegreeW. [very] [hungry]	Deg-A	Deg-A	Deg-A	A-Deg	Deg-A	Deg-A	Deg-A

Table 4
The distribution of main word order types (Subject, Object, Verb) in the world languages. From the World Atlas of Language Structures, chapter 81 (Dryer 2011).

Order	**Languages**	
SOV	565	41%
SVO	488	35%
VSO	95	7%
VOS	25	2%
OVS	11	1%
OSV	4	<1%
mixed/no-dominant	189	14%
total sample size	1,377	

4.2.3 Language Sample. For our study, we have chosen seven widely spoken languages. These are English, German, French, Arabic (Modern Standard), Turkish, Japanese, and Chinese (Mandarin). Mainly based on the WALS, we have summarized the word order feature values for all these languages in Table 3. Whenever possible, features were assigned one (or two) values corresponding to the dominant order(s) in that language. When no particular order was given as dominant we marked it as "mixed."

The main word order of German and Arabic deserves a special mention. In German, the positioning of subject, object, and verb is syntactically determined: main clauses without auxiliary verb are SVO, while subordinate clauses and clauses containing an auxiliary are SOV. A further complication, not marked in Table 3, is that the German finite verb must be placed in second position, in which case the pattern becomes S*Aux*OV, with the object intervening between auxiliary and main verb. As regards Arabic, whereas the WALS classifies Modern Standard Arabic as VSO, the corpora typically used in SMT show a very mixed distribution of VSO and SVO clauses.[16] Carpuat, Marton, and Habash (2012) examined the Arabic–English Treebank and found that, when the subject is expressed, it follows the verb in 70% of the cases, but precedes it in 30%. Similarly, in the Pennsylvania Arabic Treebank, they found an order distribution of 67% VS and 33% SV. Besides frequency, it can be noted that the SVO sentences attested in these corpora are in general pragmatically neutral. We conjecture that this variability in Modern Standard Arabic may be due to the effect of spoken language varieties such as Egyptian, Gulf, Kuwaiti, Iraqi (all listed as SVO by the WALS), and Syrian (listed as VSO/SVO). For these reasons, we classify Arabic as a mixed VSO/SVO language.

It is worth noting that our seven-language sample covers the main word order types of the large majority of the world languages: namely, SOV, SVO, and VSO (see Table 4).

4.3 Word Order Differences

Linguistically motivated word order profiles can be very helpful anticipating the kind of word reordering problems that an SMT system will have to face. Clearly, these will also vary in relation to the text genre (written news, speeches, etc.) and to the translation's style and degree of literality. However, we can reasonably expect the syntactic properties of the two languages to determine the general reordering characteristics of that pair.

We will now analyze the reordering characteristics of seven language pairs: English paired with the other six languages presented in Table 3, as well as the French and Arabic pair. To this end, we propose the following analysis procedure. As a first indication of reordering complexity, we look at the main word order feature (subject, object, verb). A difference at this level typically results in poor SMT performances. Then, we count the total number of discordant features. To simplify, if a particular element does not exist in a language (e.g., polar question particles in English) we count *zero* difference for that feature, and when one of the languages has a mixed order we count a *half* difference. We insist, however, on the qualitative nature of our analysis: Numbers are only meaningful in combination with the list of specific discordant features, as these have a different impact on word reordering. In particular, we find it essential for SMT to distinguish between clause-level and phrase-level differences (**CDiff** and **PDiff**) because the former account for most longer-range word movements, and the latter for the shorter. Thus, a language pair with only phrase-level discordant features is

[16] VOS order is also admitted in Arabic, but only in specific contexts (e.g., when the object is expressed by a pronoun).

likely to be suitable for a PSMT approach, where reordering is managed through local distortion or inside translation units. On the contrary, the presence of many clause-level differences typically calls for a tree-based solution, either at preprocessing or at decoding time. As we will see, some pairs lie on the borderline, with only one or few clause-level differences. Finally, it should be noted that, even among features of the same group, some have more impact on SMT than others due to their frequency or to the average length of their constituents. For instance, the order of noun and genitive is more important than that of adjective and degree word.

English and German [Main order: different; CDiff: 1.5; PDiff: 0.5]
The main word order of German is SVO or SOV, according to the syntactic context (cf. Section 4.2). German also differs from English with respect to the position of oblique phrases and that of the negation: Both are fixed in English but mixed in German. At the phrase level, German predominantly places the genitive after the noun, while English displays both orders.

Thus, despite belonging to the same family branch (Indo-European/Germanic), this pair displays complex reordering patterns. Indeed, German–English reordering has been widely studied in SMT and is still an open topic. At the Workshop on Statistical Machine Translation 2014 (Bojar et al. 2014), a syntax-based string-to-tree SMT approach (Williams et al. 2014) won in both language directions (official results excluding online systems). At the International Workshop on Spoken Language Translation 2014 (Cettolo et al. 2014), the best submission was a combination of PSMT with POS- and syntax-based preordering (Slawik et al. 2014), string-to-tree syntax-based SMT, and factored PSMT (Birch et al. 2014).

English and French [Main order: same; CDiff: 0.5; PDiff: 1.5]
Most clause-level features have the same values in French as in English, except for the negation, which is typically expressed by two words in French: one preceding and one following the verb.[17] At the phrase level, differences are found in the location of genitives and adjectives. Thus, English and French have very similar clause-level orders, but reordering is abundant at the local level.

This is a case where reordering is mostly well handled by string-based PSMT. As a reference, the three top English-to-French WMT14 systems (official results excluding online systems) were all phrase-based. A similar trend was observed in the French-to-English track.

English and Arabic [Main order: different; CDiff: 0.5; PDiff: 2.5]
The dominant Arabic order is VSO, followed by SVO (cf. Section 4.2). Apart from this important difference, all other clause-level features agree between Arabic and English. At the phrase level, differences are found in genitives, adjectives, and degree words.

As a result, reordering is overwhelmingly local but few crucial long-range reorderings also regularly occur. Thus, this pair is challenging for PSMT but, at the same time, not well suited for a tree-based approach. As shown by Zollmann et al. (2008) and Birch, Blunsom, and Osborne (2009), PSMT performs similarly or better than HSMT for the Arabic-to-English language pair. However, HSMT was shown to better cope with the reordering of VSO sentences (Bisazza 2013). Pre-ordering of Arabic VSO sentences for translation into English has proved to

[17] Pre-verbal negation can be omitted in colloquial French.

be a particularly difficult task (Green, Sathi, and Manning 2009; Carpuat, Marton, and Habash 2010) and has inspired work on hybrid pre-ordering where multiple verb pre-orderings are fed to a PSMT decoder (Bisazza and Federico 2010; Andreas, Habash, and Rambow 2011); see also Section 2.4.

English and Turkish [Main order: different; CDiff: 5.5; PDiff: 1.5]
Turkish is a good example of head-final language, except for the fact that it can use both clause-final and clause-initial subordinators.[18] As a result, almost all clause-level features are discordant in this pair. At the phrase level, Turkish mainly differs from English for the use of postpositions instead of prepositions.

Among our language pairs, this is one of the most difficult to reorder for an SMT system. The complex nature of its reordering phenomena suggests a good fit for tree-based SMT approaches, and indeed, HSMT was shown to significantly outperform PSMT between Turkish and English in both language directions (Ruiz et al. 2012; Yılmaz et al. 2013). However, state-of-the-art SMT quality in this language pair is still very low, mostly because of the agglutinative nature of Turkish, which makes it difficult to tear apart word reordering issues from rich morphology issues. Attempting to address both issues in an English-to-Turkish factored PSMT system, Yeniterzi and Oflazer (2010) pre-process the parsed English side with a number of syntax-to-morphology mapping rules and constituent pre-ordering rules dealing with local and global reordering phenomena, respectively. Only the former, though, resulted in better translation quality.

English and Japanese [Main order: different; CDiff: 6; PDiff: 1.5]
Japanese is the prototypical example of head-final language. In this pair all clause-level features are discordant, whereas at the phrase level, Japanese differs from English for the use of postpositions and the strictly head-final genitive construction. This pair, like the previous one, is extremely challenging for PSMT because of the hierarchical nature of its reordering phenomena and the high frequency of long-range word movements. Indeed, translation between English and Japanese has spurred a remarkable amount of work on pre-ordering, post-ordering, and decoding-time reordering. In 2013 the PatentMT evaluation campaign of the NTCIR conference (Goto et al. 2013a) saw rule-based and hybrid systems largely outperform the purely statistical ones in Japanese-to-English. The highest-ranked SMT submission was actually a combination of three SMT systems, including a baseline PSMT method, a rule-based pre-ordering method, and a post-ordering method based on string-to-tree syntax-based SMT (Sudoh et al. 2013). Interestingly, the trends were different in the opposite translation direction, English-to-Japanese, where all rule-based MT systems were significantly outperformed by a PSMT system that performed pre-ordering of the English input with few manual rules for head finalization based on dependency parse trees (Sudoh et al. 2013).

English and Chinese [Main order: same; CDiff: 3.5; PDiff: 1]
Despite belonging to the same main order type, these two languages differ in the positioning of oblique phrases, relative clauses, interrogative phrases, and

[18] In Turkish, non-finite subordinate clauses are typically placed before the main clause and linked to it by a clause-final subordinator (e.g., *rağmen/although*), whereas finite subordinate clauses can be placed after the main clause and introduced by a clause-initial subordinator (e.g., *ama/but*). The former is dominant in written language.

subordinating words.[19] Moreover, word order variations are quite common in Chinese to mark the *topic* of a sentence, (i.e., what is being talked about). Comparing the two languages at the phrase level, we find partial disagreement in the use of genitive and adpositions (Chinese has both prepositions and postpositions).

Thus, this pair too is characterized by very complex reordering, hardly manageable by a PSMT system. This is confirmed by a number of empirical results showing that tree-based approaches (particularly HSMT) consistently outperform PSMT in Chinese-to-English evaluations (Zollmann et al. 2008; Birch, Blunsom, and Osborne 2009). It is worth noting that translation between Chinese and English has been the main motivation and test bed for the development of HSMT.

French and Arabic [Main order: different; CDiff: 1.5; PDiff: 1]

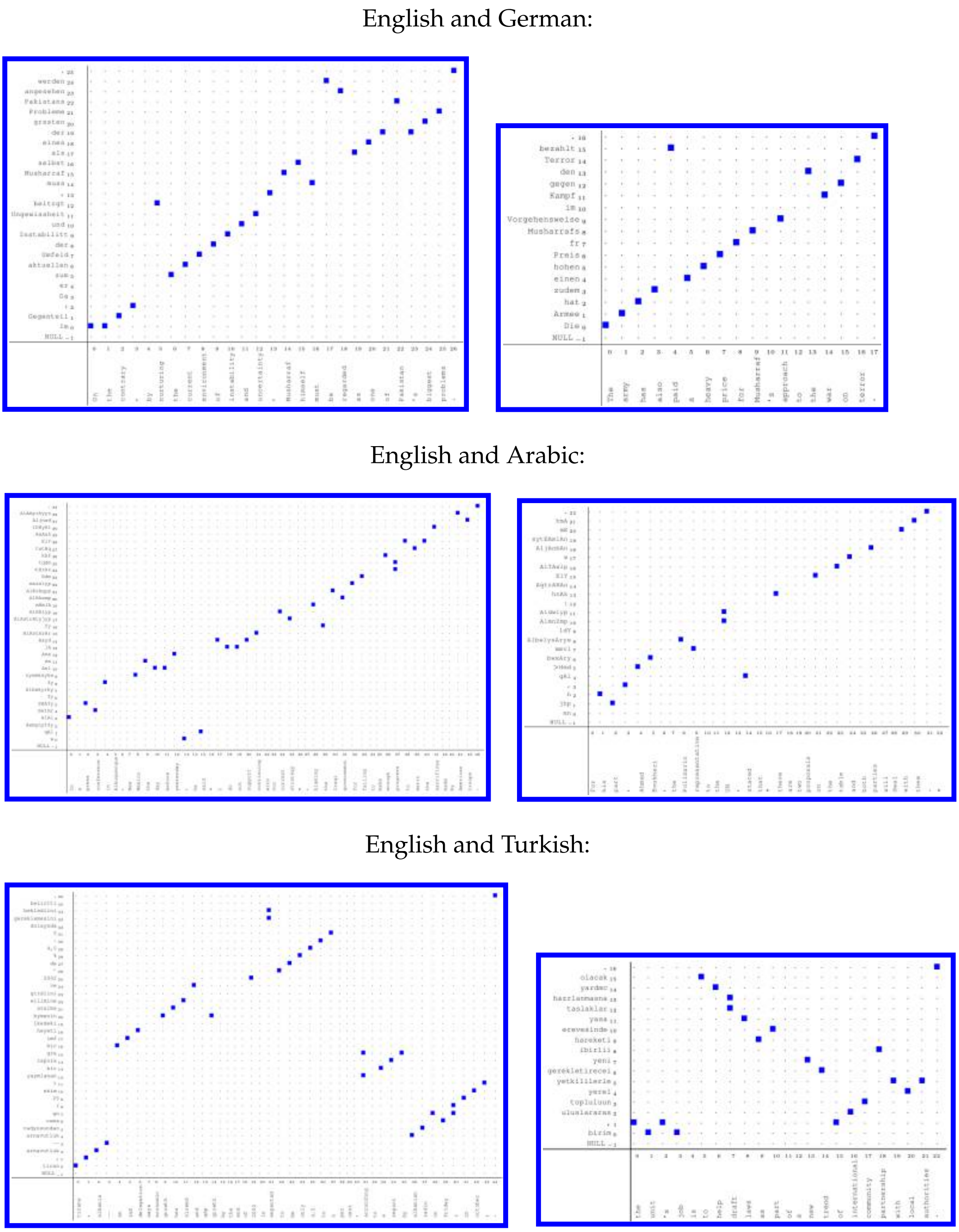
At the clause level, this pair differs in main word order (SVO versus VSO or SVO) like the English–Arabic pair, but also in the order of negation and verb. On the other hand, phrase-level order is notably more similar, with only one discordant feature of minor importance (adjective and degree word).

Less research was published on this language pair. Nevertheless, Hasan and Ney (2008) and Schwenk and Senellart (2009) chose a PSMT approach to experiment with an Arabic-to-French task.

Figure 8 illustrates the reordering characteristics of three language pairs by means of sentence examples that were automatically word-aligned with GIZA++ (Och and Ney 2003) (intersection of direct and inverse alignments). In the first row, we see two English–German sentence pairs; in both cases, most of the points lie close to the diagonal representing an overall monotonic translation, whereas few isolated points denote the very long-range reordering of verbs. Similarly, in the two English–Arabic sentence pairs, we mostly observe local reorderings, with the exception of few isolated points corresponding to the Arabic clause-initial verbs. Finally, the two Turkish–English examples display global reordering, due to the high number of clause-level order differences.

Where possible, it is interesting to relate our analysis with previously published measures of reordering based on parallel data. To our knowledge, the most comprehensive results of this kind are reported by Birch (2011), who formulates reordering as a binary process occurring between two blocks that are adjacent in the source (cf. ITG constraints in Section 2.1). Here, the general amount of reordering in a language pair is estimated by the RQuantity, defined as the sum of the spans of all the reordered blocks on the target side, normalized by the length of the target sentence and averaged over a corpus. Based on the Europarl corpus (Koehn 2002) and automatic word alignments, Birch (2011) reports average RQuantity values of 0.586/0.608 in English-to-German/German-to-English, versus only 0.402/0.395 in English-to-French/French-to-English. The manually aligned GALE corpus (LDC2006E93) is instead used to measure the distribution of reordering widths, defined as the sum of the swapped blocks' target spans. Widths are binned into short (2–4 words), medium (5–8), and long (>8). In Chinese-to-English there are about 0.8/0.9/0.9 short/medium/long reordered blocks per sentence, whereas in Arabic-to-English there are 1.1/0.4/0.2 short/medium/long reordered blocks per sentence. These figures align nicely with our classification of phrase- and clause-level differences, which we have related to longer and shorter-range

[19] Subordinating words in Chinese can occur at the beginning of the subordinate clause, at its end, or even inside it (Li 2008).



English and German:

English and Arabic:

English and Turkish:

Figure 8
Word-alignment matrices of sentence pairs taken from three parallel news corpora: the NIST-MT-08 Arabic-English evaluation benchmark, the WMT-10 German-English training corpus, and the Turkish-English South European Times corpus (Tyers and Alperen 2010). English is always on the *x*-axis.

reordering, respectively: Chinese-to-English (PDiff: 1, CDiff: 3.5) displays much more reordering overall, whereas Arabic-to-English (PDiff: 2.5, CDiff: 0.5) has more short reorderings but much fewer medium and short.

The advantage of using our proposed analysis is that it can be easily extended to other language pairs thanks to the wide coverage of WALS, whereas data-driven analyses depend on the availability of high-quality word-aligned parallel corpora.

5. Discussion and Conclusions

We have provided a comprehensive overview of how the word reordering problem is modeled within different string-based and tree-based SMT frameworks, and as a stand-alone task. To summarize, string-based SMT considers all permutations of the source sentence and relies on separate reordering models to score them. On the other hand, tree-based SMT tightly couples reordering to translation and, during decoding, only or mostly considers word permutations that are licensed by the learned translation model. In practice, both approaches apply general heuristic constraints on the maximum reordering width to avoid explosion of the search space.

The main weakness of a string-based approach like phrase-based SMT (PSMT) with regard to reordering lies in its coarse definition of the reordering search space. In this framework, relaxing the distortion limit means dramatically increasing the size of the search space, making the reordering model's task extremely complex and intensifying the risk of both search and model errors. As a result, PSMT is generally good at handling local reordering but largely fails to capture long-range reordering phenomena.

As for tree-based SMT, a distinction must be made between methods that extract hierarchical structure directly from parallel data and methods that rely on syntactic annotation provided by pre-trained monolingual parsers. A prominent example of the former is hierarchical phrase-based SMT (HSMT), which models reordering via partially lexicalized translation rules. Although this results in a more principled definition of the reordering search space, HSMT lacks the ability to generalize the learned reordering patterns from specific lexical clues to whole word or phrase categories.

Finally, reordering may be constrained by syntactic information in the source or target language, or both. When syntax is used in the source language, reordering is performed by transforming a given parse tree of the input sentence. When syntax is used in the target language, reordering is allowed only if resulting in a grammatically valid target tree fragment. Syntactic information is adopted by both syntax-based SMT, where the tree is reordered and translated simultaneously, and by syntactic pre-ordering (or post-ordering) methods, where the tree is transformed before (or after) translation. The success of these approaches largely depends on the degree of isomorphism of the modeled language pair, as well as on the parser's performance, which can vary substantially across languages.

After describing how word reordering is modeled in SMT, we have questioned why different language pairs appear to need different reordering modeling solutions. To answer this question, we have outlined the word order profiles of seven widely spoken languages, based on a large body of linguistic knowledge. Then we have examined their pairwise differences in detail. Finally, we have used these differences to interpret the empirical findings of previous work that evaluated various SMT reordering techniques in those language pairs.

We conclude from our analysis that a few linguistic facts can be very useful to predict the reordering characteristics of a language pair and to select the SMT approach that best suits them. In particular, string-based PSMT is preferable for language pairs with only constituent-level differences, like French–English, as these mostly imply short- or medium-range reordering patterns that can be captured by local distortion. On the other hand, language pairs with many clause-level order differences (e.g., Japanese–English, Turkish–English, Chinese–English) are best handled by tree-based SMT or syntax-based pre-/post-ordering approaches that can handle complex, hierarchical reordering patterns. While this may seem obvious, we notice that, in the literature, the choice of an optimal SMT framework for a new translation task is often driven by costly empirical trials rather than by linguistic knowledge. Finally, the pairs with mostly constituent-level differences and only one or few clause-level differences (e.g., German–English and Arabic–English) do not fit well into either category. In sentences without global reordering, HSMT can underperform PSMT, likely because of the much larger search space explored. At the same time, applying PSMT to such pairs with heuristic reordering constraints can lead to systematic errors in the positioning of important elements of the sentence, such as verbs. Not surprisingly, these language pairs have been the object of a fair amount of work aimed at refining the reordering space of both PSMT and HSMT. Our word order analysis can be easily extended to other language pairs, using the methodology presented in Section 4.

In conclusion, finding a definitive solution to the problem of word reordering implies answering the fundamental research questions of SMT: Is structure needed to translate? If so, what kind of structure and how should it be used? A growing part of the research community has converged on a positive answer to the former question, but the latter remains open to date. While the field keeps evolving around these questions, SMT has already reached the stage of applied language technology. We hope this survey will provide practical guidelines to the system developers of today and, at the same time, good scientific references to the researchers elaborating the solutions of tomorrow.

Acknowledgments

We would like to thank Alexandra Birch, Marta R. Costa-jussà, Nadir Durrani, Chris Dyer, Adrià de Gispert, Isao Goto, Spence Green, Zhongqiang Huang, Maxim Khalilov, Graham Neubig, Khalil Sima'an, Miloš Stanojević, Katsuhito Sudoh, Christoph Tillmann, Taro Watanabe, and Richard Zens, as well as the anonymous reviewers, for providing valuable feedback on an earlier version of this survey.

References

Al-Onaizan, Yaser and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536, Sydney.

Andreas, Jacob, Nizar Habash, and Owen Rambow. 2011. Fuzzy syntactic reordering for phrase-based statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 227–236, Edinburgh.

Auli, Michael, Michel Galley, and Jianfeng Gao. 2014. Large-scale expected BLEU training of phrase-based reordering models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1250–1260, Doha.

Bach, Nguyen, Stephan Vogel, and Colin Cherry. 2009. Cohesive constraints in a beam search phrase-based decoder. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 1–4, Boulder, CO.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI.

Bangalore, Srinivas and Giuseppe Riccardi. 2000. Finite-state models for lexical reordering in spoken language translation. In *Proceedings of International Conference on*

Spoken Language Processing, volume 2, pages 422–425, Beijing.

Berger, Adam L., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. Kehler, and Robert L. Mercer. 1996. Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981, Apr.

Bikel, Daniel M. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Birch, Alexandra. 2011. *Reordering Metrics for Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Birch, Alexandra, Phil Blunsom, and Miles Osborne. 2009. A quantitative analysis of reordering phenomena. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Morristown, NJ.

Birch, Alexandra, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, and Philipp Koehn. 2014. Edinburgh SLT and MT system description for the IWSLT 2014 evaluation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 49–56, Lake Tahoe, CA.

Birch, Alexandra and Miles Osborne. 2011. Reordering metrics for MT. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1027–1035, Portland, OR.

Birch, Alexandra, Miles Osborne, and Phil Blunsom. 2010. Metrics for MT evaluation: Evaluating reordering. *Machine Translation*, 24(1):15–26.

Birch, Alexandra, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Stroudsburg, PA.

Bisazza, Arianna. 2013. *Linguistically Motivated Reordering Modeling for Phrase-Based Statistical Machine Translation*. Ph.D. thesis, University of Trento.

Bisazza, Arianna and Marcello Federico. 2010. Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 241–249, Uppsala.

Bisazza, Arianna and Marcello Federico. 2012. Modified distortion matrices for phrase-based statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–487, Jeju Island.

Bisazza, Arianna and Marcello Federico. 2013a. Dynamically shaping the reordering search space of phrase-based statistical machine translation. *Transactions of the ACL*, 1:327–340.

Bisazza, Arianna and Marcello Federico. 2013b. Efficient solutions for word reordering in German-English phrase-based statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 440–451, Sofia.

Bisazza, Arianna, Daniele Pighin, and Marcello Federico. 2012. Chunk-lattices for verb reordering in Arabic-English statistical machine translation. *Machine Translation, Special Issue on MT for Arabic*, 26(1-2):85–103.

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD.

Braune, Fabienne, Anita Gojun, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based SMT. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–30, Trento.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

Carpuat, Marine, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English SMT by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183, Uppsala.

Carpuat, Marine, Yuval Marton, and Nizar Habash. 2012. Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment. *Machine Translation, Special Issue on MT for Arabic*, 26(1-2):105–120.

Casacuberta, Francisco and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 2–17, Lake Tahoe, CA.

Chang, Pi-Chuan and Kristina Toutanova. 2007. A discriminative syntactic word order model for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Prague.

Chen, Boxing, Mauro Cettolo, and Marcello Federico. 2006. Reordering rules for phrase-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 182–189, Kyoto.

Cherry, Colin. 2008. Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 72–80, Columbus, OH.

Cherry, Colin. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Atlanta, GA.

Cherry, Colin, Robert C. Moore, and Chris Quirk. 2012. On hierarchical re-ordering and permutation parsing for phrase-based decoding. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 200–209, Montréal.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, MI.

Chiang, David, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 224–233, Honolulu, HI.

Collins, Michael, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, MI.

Corder, Stephen Pit. 1979. Language distance and the magnitude of the language learning task. *Studies in Second Language Acquisition*, 2(01):27–36.

Costa-jussà, Marta R. and José A. R. Fonollosa. 2006. Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 70–76, Sydney.

Costa-jussà, Marta R. and José A. R. Fonollosa. 2009. State-of-the-art word reordering approaches in statistical machine translation: A survey. *IEICE TRANSACTIONS on Information and Systems*, E92-D(11):2179–2185.

Crego, Josep M. and Nizar Habash. 2008. Using shallow syntax information to improve word alignment and reordering for SMT. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61, Morristown, NJ.

Crego, Josep M., José B. Mariño, and Adrià de Gispert. 2005. Reordered search, and tuple unfolding for ngram-based SMT. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 283–289, Phuket.

Crego, Josep Maria and José B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

de Gispert, Adrià, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. *Computational Linguistics*, 36(3):505–533.

DeNero, John and Jakob Uszkoreit. 2011. Inducing sentence structure from parallel corpora for reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 193–203, Stroudsburg, PA.

Diab, Mona, Kadri Hacioglu, and Daniel Jurafsky. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 149–152, Boston, MA.

Ding, Yuan and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual*

Meeting of the Association for Computational Linguistics (ACL'05), pages 541–548, Ann Arbor, MI.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language*, 13:257–292.

Dryer, Matthew S. 2007. Word order. In Timothy Shopen, editor, *Clause Structure, Language Typology and Syntactic Description*, volume 1. Cambridge University Press, second edition, chapter 2, pages 61–131.

Dryer, Matthew S. 2011. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Dryer, Matthew S. and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011 edition.

Durgar El-Kahlout, İlknur and Kemal Oflazer. 2010. Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1313–1322.

Durrani, Nadir, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can Markov models over minimal translation units help phrase-based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Sofia.

Durrani, Nadir, Philipp Koehn, Helmut Schmid, and Alexander Fraser. 2014. Investigating the usefulness of generalized word representations in SMT. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 421–432, Dublin.

Durrani, Nadir, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, OR.

Dyer, Chris and Philip Resnik. 2010. Context-free reordering, finite-state translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 858–866, Los Angeles, CA.

Elming, Jakob and Nizar Habash. 2009. Syntactic reordering for English-Arabic phrase-based machine translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 69–77, Athens.

Feng, Minwei, Arne Mauser, and Hermann Ney. 2010. A source-side decoding sequence model for statistical machine translation. In *Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, CO.

Feng, Minwei, Jan-Thorsten Peter, and Hermann Ney. 2013. Advancements in reordering models for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 322–332, Sofia.

Feng, Minwei, Weiwei Sun, and Hermann Ney. 2012. Semantic cohesion model for phrase-based SMT. In *Proceedings of COLING 2012*, pages 867–878, Mumbai.

Feng, Yang, Haitao Mi, Yang Liu, and Qun Liu. 2010. An efficient shift-reduce decoding algorithm for phrased-based machine translation. In *COLING (Posters)*, pages 285–293, Beijing.

Finch, Andrew and Eiichiro Sumita. 2009. Bidirectional phrase-based statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1132, Singapore.

Fox, Heidi. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 304–311, Philadelphia, PA.

Freitag, Markus, Minwei Feng, Matthias Huck, Stephan Peitz, and Hermann Ney. 2013. Reverse word order models. In *Machine Translation Summit*, pages 159–166, Nice.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 273–280, Boston, MA.

Galley, Michel and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Morristown, NJ.

Gao, Yang, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 857–868, Edinburgh.

Genzel, Dmitriy. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 376–384, Stroudsburg, PA.

Gojun, Anita and Alexander Fraser. 2012. Determining the placement of German verbs in English-to-German SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 726–735, Avignon.

Goto, Isao, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of NTCIR-10*, pages 260–286, Tokyo.

Goto, Isao, Masao Utiyama, and Eiichiro Sumita. 2012. Post-ordering by parsing for Japanese-English statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 311–316, Jeju Island.

Goto, Isao, Masao Utiyama, and Eiichiro Sumita. 2013. Post-ordering by parsing with ITG for Japanese-English statistical machine translation. *ACM Transactions on Asian Language Information Processing*, 12(4):17:1–17:22.

Goto, Isao, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 155–165, Sofia.

Green, Spence, Michel Galley, and Christopher D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 867–875, Los Angeles, CA.

Green, Spence, Conal Sathi, and Christopher D. Manning. 2009. NP subject detection in verb-initial Arabic clauses. In *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3)*, Ottawa.

Gupta, Deepa, Mauro Cettolo, and Marcello Federico. 2007. POS-based reordering models for statistical machine translation. In *Proceedings of MT Summit XI*, pages 207–213, Copenhagen.

Habash, Nizar. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of the Machine Translation Summit XI*, pages 215–222, Copenhagen.

Hardmeier, Christian, Arianna Bisazza, and Marcello Federico. 2010. FBK at WMT 2010: Word lattices for morphological reduction and chunk-based reordering. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 88–92, Uppsala.

Hasan, Saša and Hermann Ney. 2008. A multi-genre SMT system for Arabic to French. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, pages 2167–2170, Marrakech.

Hayashi, Katsuhiko, Katsuhito Sudoh, Hajime Tsukada, Jun Suzuki, and Masaaki Nagata. 2013. Shift-reduce word reordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1382–1386, Seattle, WA.

He, Zhongjun, Yao Meng, and Hao Yu. 2010. Extending the hierachical phrase based model with maximum entropy based BTG. In *Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, CO.

Hopkins, Mark and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh.

Howlett, Susan and Mark Dras. 2011. Clause restructuring for SMT not absolutely helpful. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 384–388, Portland, OR.

Huang, Liang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 66–73, Cambridge, MA.

Huang, Zhongqiang, Jacob Devlin, and Rabih Zbib. 2013. Factored soft source syntactic constraints for hierarchical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 556–566, Seattle, WA.

Huck, Matthias, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A phrase orientation model for hierarchical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 452–463, Sofia.

Imamura, Kenji, Hideo Okuma, and Eiichiro Sumita. 2005. Practical approach to syntax-based statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 267–274, Phuket.

Isozaki, Hideki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA.

Isozaki, Hideki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 244–251, Uppsala.

Jehl, Laura, Adrià de Gispert, Mark Hopkins, and Bill Byrne. 2014. Source-side preordering for translation using logistic regression and depth-first branch-and-bound search. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 239–248, Gothenburg.

Kanthak, Stephan, David Vilar, Evgeny Matusov, Richard Zens, and Hermann Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174, Ann Arbor, MI.

Katz-Brown, Jason, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. 2011. Training a parser for machine translation reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 183–192, Edinburgh.

Khalilov, Maxim and José A. R. Fonollosa. 2011. Syntax-based reordering for statistical machine translation. *Computer Speech and Language*, 25:761–788.

Khalilov, Maxim and Khalil Sima'an. 2011. Context-sensitive syntactic source-reordering by statistical transduction. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 38–46, Chiang Mai.

Khalilov, Maxim and Khalil Sima'an. 2012. Statistical translation after source reordering: Oracles, context-aware models, and empirical analysis. *Natural Language Engineering*, 18(04):491–519.

Klein, Dan and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS)*. MIT Press, Cambridge, MA, pages 3–10.

Knight, Kevin. 1999. Decoding complexity in word replacement translation models. *Computational Linguistics*, 25(4):607–615.

Koehn, Philipp. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, `http://www.isi.edu/∼koehn/europarl/`.

Koehn, Philipp, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, PA.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133, Edmonton.

Lerner, Uri and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 513–523, Seattle, WA.

Li, Chi-Ho, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 720–727, Prague.

Li, Junhui, Zhaopeng Tu, Guodong Zhou, and Josef van Genabith. 2012. Using syntactic head information in hierarchical phrase-based translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 232–242, Montréal.

Li, Peng, Yang Liu, and Maosong Sun. 2013. Recursive autoencoders for ITG-based translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 567–577, Seattle, WA.

Li, Peng, Yang Liu, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2014. A neural reordering model for phrase-based translation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1897–1907, Dublin.

Li, Ying. 2008. Three sensitive positions and Chinese complex sentences: A comparative perspective. *Journal of Chinese Language and Computing*, 18(2):47–59.

Liu, Yang, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 609–616, Stroudsburg, PA.

Maillette de Buy Wenniger, Gideon and Khalil Sima’an. 2014. Bilingual Markov reordering labels for hierarchical SMT. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–21, Doha.

Marcu, Daniel, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney.

Mariño, J. B., R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-Jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Marton, Yuval and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-08: HLT*, pages 1003–1011, Columbus, OH.

Menezes, Arul and Chris Quirk. 2007. Using dependency order templates to improve generality in translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 1–8, Prague.

Moore, Robert C. and Chris Quirk. 2007. Faster beam-search decoding for phrasal statistical machine translation. In *Proceedings of MT Summit XI*, pages 321–327, Copenhagen.

Mylonakis, Markos and Khalil Sima’an. 2010. Learning probabilistic synchronous CFGS for phrase-based translation. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 117–125, Uppsala.

Mylonakis, Markos and Khalil Sima’an. 2011. Learning hierarchical translation structure with linguistic annotations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 642–652, Portland, OR.

Nagata, Masaaki, Kuniko Saito, Kazuhide Yamamoto, and Kazuteru Ohashi. 2006. A clustered global phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 713–720, Sydney.

Neubig, Graham, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853, Jeju Island.

Nguyen, ThuyLinh and Stephan Vogel. 2013. Integrating phrase-based reordering features into a chart-based decoder for machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1587–1596, Sofia.

Niehues, Jan and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 206–214, Athens.

Nießen, Sonja and Hermann Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of the MT Summit VIII: Machine Translation in the Information Age*, pages 247–252, Santiago de Compostela.

Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC-2006*, pages 2216–2219, Genoa.

Och, Franz Josef. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for*

Computational Linguistics (EACL), pages 71–76, Bergen.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Research Report RC22176, IBM Research Division, Thomas J. Watson Research Center.

Popović, Maja and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genoa.

Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, MI.

Rottmann, Kay and Stephan Vogel. 2007. Word reordering in statistical machine translation with a POS-based distortion model. In *Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 171–180, Skövde.

Ruiz, Nick, Arianna Bisazza, Roldano Cattoni, and Marcello Federico. 2012. FBK's Machine Translation Systems for IWSLT 2012's TED Lectures. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 61–68, Hong Kong.

Schwenk, Holger and Jean Senellart. 2009. Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *Proceedings of the Machine Translation Summit XII*, Ottawa.

Setiawan, Hendra, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 712–719, Prague.

Setiawan, Hendra, Min Yen Kan, Haizhou Li, and Philip Resnik. 2009. Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 324–332, Suntec.

Setiawan, Hendra, Bowen Zhou, and Bing Xiang. 2013. Anchor Graph: Global reordering contexts for statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 501–512, Seattle, WA.

Setiawan, Hendra, Bowen Zhou, Bing Xiang, and Libin Shen. 2013. Two-neighbor orientation model with cross-boundary global contexts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Sofia.

Shen, Libin, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency statistical machine translation. *Computational Linguistics*, 36(4):649–671.

Slawik, Isabel, Mohammed Mediani, Jan Niehues, Yuqi Zhang, Eunah Cho, Teresa Herrmann, Thanh-Le Ha, and Alex Waibel. 2014. The KIT translation systems for IWSLT 2014. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 119–126, Lake Tahoe, CA.

Smith, David and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 23–30, New York, NY.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh.

Stanojević, Miloš and Khalil Sima'an. 2014a. Evaluating word order recursively over permutation-forests. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 138–147, Doha.

Stanojević, Miloš and Khalil Sima'an. 2014b. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha.

Sudoh, Katsuhito, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. 2010. Divide and translate: Improving long distance reordering in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 418–427, Uppsala.

Sudoh, Katsuhito, Jun Suzuki, Hajime Tsukada, Masaaki Nagata, Sho Hoshino, and Yusuke Miyao. 2013. NTT-NII statistical machine translation for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, pages 294–300, Tokyo.

Sudoh, Katsuhito, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Post-ordering in Statistical Machine Translation. In *MT Summit XIII: The Thirteenth Machine Translation Summit*, pages 316–323, Xiamen.

Sudoh, Katsuhito, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2013. Syntax-based post-ordering for efficient Japanese-to-English translation. *ACM Transactions on Asian Language Information Processing*, 12(3):12:1–12:15.

Talbot, David, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz Och. 2011. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21, Edinburgh.

Tillmann, Christoph. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 101–104, Boston, MA.

Tromble, Roy and Jason Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1007–1016, Singapore.

Tyers, Francis M. and Murat Serdar Alperen. 2010. South-east European times: A parallel corpus of Balkan languages. In *Proceedings of LREC 2010, Seventh International Conference on Language Resources and Evaluation*, pages 49–53, Valletta.

Visweswariah, Karthik, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. 2011. A word reordering model for improved machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 486–496, Edinburgh.

Wang, Chao, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague.

Watanabe, Taro and Eiichiro Sumita. 2002. Bidirectional decoding for statistical machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1079–1085, Taipei.

Watanabe, Taro, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 777–784, Sydney.

Wellington, Benjamin, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 977–984, Sydney.

Williams, Philip, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's syntax-based systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, MD.

Wu, D. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1328–1335, Montreal.

Wu, Dekai. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of the 34th Annual Conference of the Association for Computational Linguistics*, pages 152–158, Santa Cruz, CA.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Xia, Fei and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, pages 508–514, Geneva.

Xiong, Deyi, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528, Sydney.

Xu, Peng, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 245–253, Boulder, CO.

Yahyaei, Sirvan and Christof Monz. 2009. Decoding by dynamic chunking for statistical machine translation. In *Proceedings of the Machine Translation Summit XII*, Ottawa.

Yahyaei, Sirvan and Christof Monz. 2010. Dynamic distortion in a discriminative reordering model for statistical machine translation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 353–360, Paris.

Yamada, Kenji and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 303–310, Philadelphia, PA.

Yang, Nan, Mu Li, Dongdong Zhang, and Nenghai Yu. 2012. A ranking-based approach to word reordering for statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–920, Jeju Island.

Yeniterzi, Reyyan and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala.

Yılmaz, Ertuğrul, İlknur Durgar El-Kahlout, Burak Aydın, Zişan Sıla Özil, and Coşkun Mermer. 2013. Tübitak Turkish-English submissions for IWSLT 2013. In *Proceedings of the International Workshop on Spoken Language Translation*, Heidelberg.

Zens, Richard. 2008. *Phrase-based Statistical Machine Translation: Models, Search, Training*. Ph.D. thesis, RWTH Aachen University.

Zens, Richard and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Sapporo.

Zens, Richard and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 55–63, New York, NY.

Zens, Richard, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of COLING 2004*, pages 205–211, Geneva.

Zens, Richard, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *25th German Conference on Artificial Intelligence (KI2002)*, pages 18–32, Aachen.

Zhang, Hao and Daniel Gildea. 2007. Factorization of synchronous context-free grammars in linear time. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 25–32, Rochester, NY.

Zhang, Min, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-08: HLT*, pages 559–567, Columbus, OH.

Zhang, Yuqi, Richard Zens, and Hermann Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop* *on Syntax and Structure in Statistical Translation*, pages 1–8, Rochester, NY.

Zollmann, Andreas and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York, NY.

Zollmann, Andreas, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 1145–1152, Manchester.