Tel Aviv University

Lester and Sally Entin Faculty of the Humanities

Department of Linguistics

Transfer-based Machine Translation between morphologically-rich and resource-poor languages: The case of Hebrew and Arabic

MA thesis submitted by

Reshef Shilon

Prepared under the guidance of

Shuly Wintner, Computer Science, University of Haifa

Fred Landman, Linguistics, Tel Aviv University

March 2011

Table of contents

1. Introduction

1.1 Motivation

Hebrew and Arabic are closely related Semitic languages. However, they are mutually incomprehensible languages with complex morphology and scarce parallel corpora. The contemporary dominant statistical Machine Translation (MT) paradigm requires large volumes of sentence-aligned parallel corpora. Unfortunately, such abundant parallel corpora currently exist only for few language pairs; and languages with low- or medium-level of availability of digitally stored material (Varga et al., 2005) require alternative approaches. Specifically, no high-quality parallel corpora exist for Hebrew–Arabic. Machine translation between the two languages is therefore interesting and challenging.

This work will detail the challenges and possible solutions to the problems that arise from translating between Hebrew and Arabic, and will discuss the solutions we implemented to these problems, with automatic evaluation scores of the output translations and manual error analysis.

1.2 Similar work

Some work has been done on MT between related languages within the same language family. An example is translation between the Turkmen and Turkish (Tantuğ et al, 2007), where the major difference between the languages is lexical and morphological (but not syntactic). In this system there was no component that mapped different syntactic structures between the languages, and the emphasis was on properly transferring and generating word-root and morphological features. Another work was done on Slavic languages (Hajic et al. 2003). The main idea was using the MT system to aid in manual translation, by using Czech as a pivot language for manual translation between English and other Slavic languages (Slovak, Polish, and Lithuanian). Only the Czech-to-Lithuanian MT system employed a module which mapped shallow syntactic structures (base phrases), and the main strategy was word-to-word lexical transfer, with no word reordering or word sense disambiguation needed. Another work was done on languages spoken in Spain, mainly Spanish, Catalan and Galician. Corbí-Bellot et al.

(2005) implemented a morphological analyzer and lexical transfer (single output word possible for an input word for a specific language pair). No complex syntactic transfer was incorporated either.

Babych et al. (2007) compared direct translation of Ukrainian-to-English to using Russian as a pivot, the latter being a language closely-related to the source language and richer in usable corpora and in resources. They concluded that in this case, using a resource-rich and linguistically-similar language as a pivot is superior to using direct translation with fewer parallel data and resources. They also showed that using a distant pivot language (in this case translating from Russian to English via French or German) does damage the output quality when sufficient resources exist for direct translation.

1.3 Work structure

Chapter 2 is an introduction to machine translation and explains useful concepts in MT. Chapter 3 discusses Stat-XFER, the framework in which we implemented our solutions. Chapter 4 surveys the main linguistic similarities and difference between Hebrew and Arabic. Chapter 5 discusses the challenges that arise from these differences in the context of MT. Chapter 6 discusses possible solutions to these challenges, and chapter 7 details our solution. Chapter 8 discusses evaluation and error analysis. Chapter 9 discusses the problem of translating prepositions as a linguistically-interesting question in the context of MT.

2. Introduction to Machine Translation

This introductory chapter surveys the history and evolution in the field of Machine Translation (MT) with its main paradigms, and introduces key concepts which will be relevant for the rest of this work.

2.1 History of MT

The History of MT starts after World War II, when the first usage of machines to decipher German encryption gave room to the idea of using machines to translate between different languages. The idea of “noisy channel” was coined, which argued that a French text is actually English seen through a noisy channel, and that the English text need only be discovered from the noisy data. As the cold war began, Russian became the main language of interest to the west, and resources were directed to that direction. The Georgetown experiment in 1954 included translation into English of about 60 Russian sentences from the domain of organic chemistry, using a rule-based system of 6 grammar rules and a vocabulary of 250 words. The experiment was regarded as highly successful, and claims were made that the problem of MT would be solved in three to five years. However, The ALPAC report in 1966, comissioned by the U.S government, argued that the ten-year long investment in MT did not live up to the expectations and claimed that more thorough research into computational linguistics was needed. This report caused massive decrease in funding, and MT research was dramatically reduced until the 1980’s. With the rise of computation power, research started again, commonly using rule-based systems such as SYSTRAN (Toma, 1977) and METEO (Chandioux, 1976).

Since the 1990's, Machine Translation (MT) has regained focus as a main research field, mainly because of the emergence of the internet and the large volumes of online text that became accessible. With it, arose statistical methods of translating texts by extracting relevant data from parallel corpora, e.g., directly translated texts in two or more languages. Brown et al. (1990, 1993) introduced Statistical Machine Translation (SMT) and IBM models 1-5. These probabilistic models were increasingly complex,

accounting for more potential correspondences between structurally different languages.

During the past decade, the statistical approaches to MT gained increasing popularity due to the abundance of available online corpora. Recently, approaches that aim at combining rich linguistic knowledge (like morphology and syntax) and statistical MT, such as factored machine translation (Koehn and Hoang, 2007) and syntax-base SMT (Yamada and Knight, 2001) are becoming more and more common.

2.2 Types of MT

The Vauquois triangle (Vauquois, 1968) is a common and useful model used for comparing translation paradigms (see figure 1 below). On one side of the ladder is the source language, and on the other side is the target language. Going up the ladder from the source sentence represents increasingly richer processing of the input sentence: from morphological analysis, through syntactic and semantic analysis, to fully disambiguated and language-independent representation of the input sentence. The phase of translating the input sentence representation into the output sentence representation is called *transfer*. If the representation of the input sentence is shallow, e.g., as a sequence of strings of characters, the process of transferring source language strings to target language string is called *direct transfer*. If the analysis of the input sentence includes syntactic or semantic processing, the process of transferring these representations is called *syntactic-transfer* or *semantic-transfer*, respectively. The final fully-disambiguated representation is called *interlingua*. The main idea is that this is a complete representation of the structure and meaning of the sentence, and it is not language-dependent.

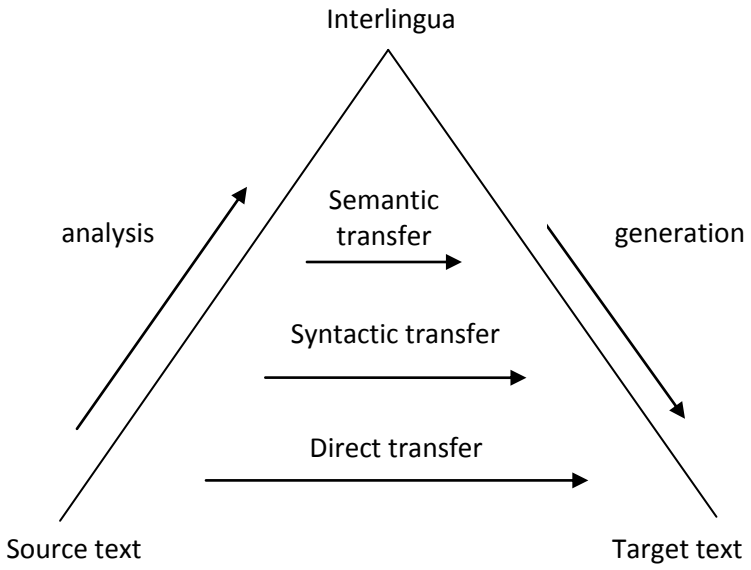

Figure 1 – Vauquois triangle of MT types

2.3 Rule-based Machine Translation (RBMT)

The classical approach to MT relies on a set of rules that map source language constructions into target language constructions, yielding a single result. Usually such systems incorporate linguistic tools such as *morphological analyzers*, which decompose a surface word into a set of morphological features, such as lemma, part-of-speech (POS), gender, number, etc. *Morphological generators* do the reverse, constructing a surface form from a set of given morpho-syntactic features. In morphologically rich languages, there is often more than one possible analysis for a given surface word, and therefore a *morphological disambiguator* can be used. This component chooses the most adequate analysis of all possible analyses in a statistical way according to the morphological and syntactic context around the word of interest. In addition, some systems also use a *syntactic parser*, which assigns a hierarchical tree structure to a given sentence, representing the syntactic relations between its constituents. Rule-based systems usually use richer representations, and therefore perform transfer "higher" in the ladder model.

The usage of such tools can create a knowledge-rich representation both of the input and the output sentences, thus allowing good means of producing fluent and grammatical output. However, it is generally accepted that these systems are less scalable, meaning that they do not perform on large-scale scenarios as well as they perform on limited vocabularies and closed domains. The limited coverage of the vocabulary and syntactic rules with the increasing ambiguity causes deterioration in result quality, and maintaining very large rule-based systems becomes very difficult. As a result, this approach is more and more regarded as obsolete.

2.4 Statistical Machine Translation (SMT)

Modern MT systems follow a common framework, which treats translation as a two-staged search problem (Brown et al. 1990, Brown et al. 1993). This framework was adopted from the field of Automatic Speech Recognition (ASR), where it proved to be efficient.

In the first stage many output sentence candidates (or *hypotheses*) in the target language are generated. These candidates are stored in a *lattice*. This stage involves a bilingual *translation lexicon* (or a *phrase table*), which gives possible translations for each word or phrase in the input sentence, with a probability assigned to every translation. Using the Vauquois terminology, this is a direct transfer.

In the second stage, a search component, called *decoder,* searches for the best output candidate according to a statistical scoring function. The output candidate is formed by concatenating consecutive hypotheses from the lattice.

Given a source language sentence S, the decoder searches for the most adequate translation in the target language. More formally, if S is a source language sentence, its best translation is the target language sentence T that is the most likely given S, that is, the one for which $\mathrm{P}(\mathrm{T}|S)$ is maximal. According to Bayes rule,

(1) $$P(T|S) = \frac{P(S|T) \cdot P(T)}{P(S)}$$

Since $\mathrm{P}(S)$ is independent of the specific translation, this amounts to a sentence $\widehat{\mathrm{T}}$ such that:

(2) $$\hat{T} = \text{argmax}_T \, P(S|T) \cdot P(T)$$

This divides the problem into a product of two probabilities, taken from two different models: a *translation model* $P(S|T)$ and a target *language model* $P(T)$.

The language model (LM) assigns a probability to each target sentence, and searches for the most fluent output sentence. The LM is usually an *n-gram model*. In this model, the sentence is divided into sequences of n tokens (typically n=3, also called *trigram*). The sentence is assigned probability by multiplying the probabilities of each of its tokens. Each token is assigned a conditional probability according to the previous two tokens. The final probability of the target sentence according to the LM is:

(3) $$P(T) = P(W_0) \cdot P(W_i|W_0) \cdot \prod_{i=2}^{n} P(W_i|W_{i-2}, W_{i-1})$$

In other words, the probability of a sentence is the product of the probabilities of the first word, the second word given the first word, and the remaining words given the two previous words. As such, this probabilistic model only looks at very local dependencies and fails to account for longer-distance dependencies (such as agreement between distant constituents).

The second part of equation (2), $P(S|T)$, is the translation probability. The decoder assigns each sentence with the probability of it being the translation of the input sentence by dividing the sentence into words (t_i and s_i) and multiplying their translation conditional probabilities. Those conditional probabilities are learned from a bilingual probabilistic lexicon.

(4) $$P(T|S) = \prod_{i=0}^{n} P(t_i|s_i)$$

In other words, the translation probability of a target sentence T given a source sentence S is the product of the translation probabilities of each target word given the source word $P(t_i|s_i)$.

This model is a raw approximation of a language and it is very easy to implement efficiently. In addition, it scales well, meaning that when moving from limited-size

systems to unrestricted larger-scale systems, the performance does not decrease and even improves. For more details, see Koehn et al. (2003).

Statistical Machine Translation uses parallel corpora to estimate word-to-word alignments and probabilities. This corpus is comprised of sentence-aligned texts in two or more languages. During translation, output candidates are generated for each source sentence according to possible translations of each input word. Koehn et al. (2003) introduced a similar kind of SMT framework, called *Phrase-Based* SMT (or PB-SMT). A phrase table is generated from aligned parallel corpora, and is used as a lexicon holding all the aligned phrases. Koehn et al. (2003) compared this model to IBM model 4 and other existing models, and showed improved results. They also argued for not taking only syntactic phrases into the phrase table, but instead allowing every aligned sequence of words to enter the phrase table. They accounted for this phenomenon with examples like ``there is'', ``with regards to'' and ``note that'', stating that although these are syntactically not constituents, they might as well be fixed phrases to be translated as units.

Some SMT systems incorporate some sort of reordering mechanism (called *distortion*) by permuting components and penalizing uncommon distortions with lower probability. However, most SMT systems do not perform any kind of syntactic or semantic analysis at all, and only perform some (local) word order changes on the output sentences. This often causes such systems to produce ungrammatical and disfluent translations.

SMT systems are also based on the existence of large parallel corpora, which only exist for a few language pairs. As a result, this framework fails to achieve quality outputs on languages with low volumes of parallel corpora.

Another challenge MT systems have to face is dealing with morphologically-rich languages. Such languages exhibit rich inflectional and derivational qualities in word forms, reflecting gender, plurality, case, tense, aspect, inflectionally-complex verb system, etc. Such languages pose a challenge to MT systems in word form analysis, word form generation, stemming, volume of parallel corpus needed, syntactic analysis complexity and language modeling.

2.5 Evaluation of MT

A topic which attracts much attention in the field of MT is automatic evaluation of system outputs. Since there are many possible correct translations for each input sentence, the task of evaluating the output is hard. The simple and naïve way is to give the outputs to human judges to score or rank. However, this is both tedious and expensive, and cannot be used to test for minor improvements over a large corpus on regular basis. Therefore, automatic metrics based on correlation to human translations are used.
First, the corpus is divided into two distinct sets – the *training* set and the *test* set. The training of the system is done exclusively on the training set, and the evaluation is done on the test set.

The most commonly-used evaluation metric is called BLEU (Papineni et al., 2002). In the common scenario, the output of the system is compared to three human translations. The score reflects correlation between the output and all the references, where correlation is measured in the amount of shared words and word-sequences between the output and human references, as well as differences in word order (also called *distortion*). This metric, despite being commonly used, has many disadvantages. First, the syntactic adequacy (like agreement or a valid parse tree) of the output is not checked. The output may therefore obtain a high score but be totally ungrammatical. Furthermore, the similarity is done on strings only, which means that if a word is mistakenly inflected in a single parameter (like plurality or case), it counts as a complete mistake. Moreover, there is no automatic account for synonymy: if an output sentence is totally correct but uses different words, it will obtain a very low score. In addition, the more human references used, the higher the score will become, regardless of the translated output.

More advanced metrics have been proposed that amend some of these flaws. METEOR (Lavie et al., 2004a) uses synonymy for matching the automatic output with the reference translations, and may also use the lemma of each word to account for partial matches between output and human reference. Other more sophisticated metrics use conservation of syntactic relations (Owczarzak, 2007) and similarity of latent semantic analysis (Reeder, 2006) when scoring the output.

3. Stat-XFER

Stat-XFER (Lavie et al., 2008) is a hybrid MT framework, which incorporates components taken from both statistical and rule-based systems. Crucially, Stat-XFER is a statistical MT framework, which uses statistical information to weigh word translations, phrase correspondences and target-language hypotheses; in contrast to other paradigms, however, it can utilize both automatically-created and manually-crafted language resources, including dictionaries, morphological processors and transfer rules. Stat-XFER has been used as a platform for developing MT systems for Hindi-to-English (Lavie et al., 2003), Hebrew-to-English (Lavie et al., 2004b), Chinese-to-English, French-to-English (Hanneman et al., 2009) and many other low-resource language pairs, such as Inupiaq-to-English or Mapudungun-to-Spanish (Monson et al., 2008).

3.1 Hybrid MT

As discussed in the previous chapter, SMT is the dominant paradigm in contemporary machine translation. However, SMT methods are problematic in handling variations in morphology and syntax between the two languages and in enforcing long-distance agreement. In addition, these methods require large volumes of parallel aligned corpora. On the other hand, traditional rule-based MT (RBMT) can handle morphology and syntax very well using a rich formalism, and can enforce agreement by using structure constraints on input and output sentences. However, these methods usually do not incorporate statistics on common or adequate translations, and it is therefore hard for such systems to scale up from a limited domain to a full free text.

The hybrid approach to MT tries to take the advantages of both frameworks while using available resources. Stat-XFER is such a hybrid MT framework, developed to specifically suit MT between morphologically-rich and resource-poor language pairs, such as Hebrew and Arabic.

In this framework, external tools can be provided and used during the process of translation. These include:

1. A Bilingual lexicon, possibly with probabilities per word-pair.
2. A Morphological analyzer of the source language

3. A Morphological disambiguator for the source language
4. A Morphological generator of the target language
5. A Statistical n-gram language model of the target language
6. A grammar which is a set of rules that map syntactic constructions from the source language to the target language. These rules may contain constraints on either language, in the format of unification-augmented SCFG rules (see below). This set of rules can either be automatically acquired from text, or manually crafted.

3.2 Context Free Grammar (CFG)

A context free grammar (CFG) is a formalism suggested by Chomsky (1956). Formally defining CFG and the process of derivation will not done in this work, and we only give an example of such a toy grammar, and a derivation tree for a short phrase which uses rules from this grammar.

```
S → NP V NP
NP → Det N
N → boy
N → girl
N → flower
Det → a
Det → the
V → likes
V → sees
```

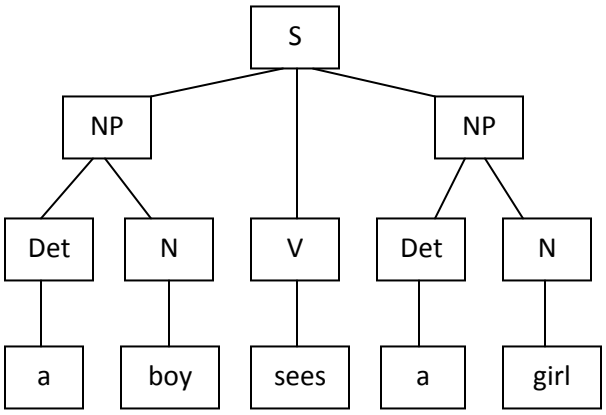

Figure 2 – An example of a context free grammar with a derivation tree of the sentence “a boy sees a girl”

3.3 Synchronous Context Free Grammar (SCFG)

SMT systems often produce ungrammatical and disfluent translation hypotheses. This is mainly because of language models that only look at local contexts (n-gram models). Syntax-based approaches to machine translation aim at fixing the problem of producing such disfluent outputs. Such syntax-based systems may take advantage of Synchronous Context Free Grammars (SCFG): mappings between two context free rules that specify how source language phrase structures correspond to target language phrase structures. For example, the following basic rule maps noun phrases in the source and target languages, and defines the order of the daughter constituents (Noun and Adjective) in both languages.

NP::NP [Det ADJ N] → [Det N ADJ]

This rule maps English NPs to French NPs, stating that an English NP is constructed from daughters Det, ADJ and N, while the French NP is constructed from daughters Det, N and ADJ, in this order. Figure 3a demonstrates translation using transfer of such derivation trees from English to French using this SCFG rule and lexical rules.

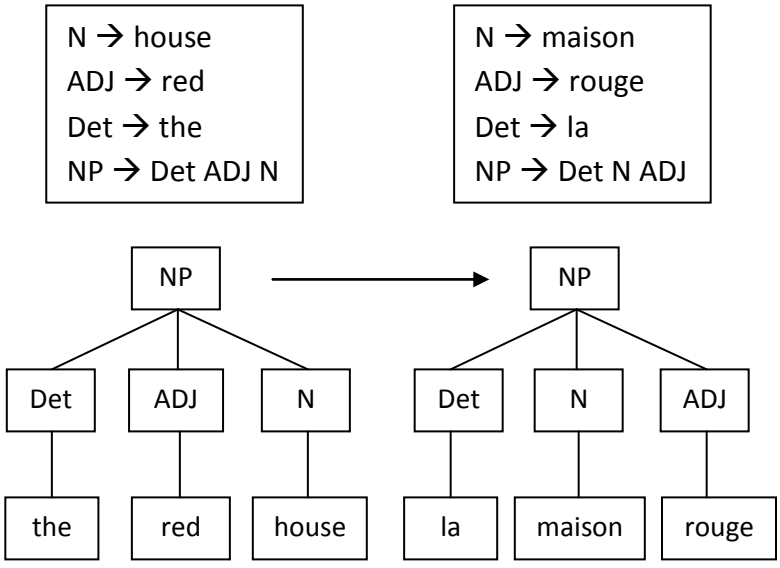

Figure 3a – translation of the English NP “The red house” into the French NP “La maison rouge” using SCFG

In SCFG translation, the source sentence is parsed according to the derivation rules, and every valid derivation of the source side leads to one or more hypotheses. The terminals (words or phrases) on the source-side are replaced with their target-side translations according to a bilingual lexicon. Syntax-based translation can still produce ungrammatical hypotheses. Since the SCFG rules typically do not cover the syntactic variety of the source language, often the system fails to produce a full parse and resorts to partial parse trees for sentence fragments. During decoding, the decoder chooses the best concatenation of these fragments based only on limited local contexts. Consequently, the resulting translations are a concatenation of target sentence fragments which were translated using SCFG rules. This may lead to ungrammatical hypotheses, e.g., because agreement is never enforced between the chosen fragments. For formal definitions and further details regarding SCFG, see Chiang (2006),

Stat-XFER uses SCFG rules to map and transfer syntactic constructions between the source language and the target language. These rules can either be automatically acquired from parallel corpora, or manually crafted.

3.4 Unification

Unification is a recursive operation done between two objects called *feature structures.* The unification matches the *value* of each matching *feature* in both feature structures, and either yields a unified feature structure, or fails altogether. The formal definition is not in the scope of this paper; for definitions and extensive introduction, see Wintner (2005), Francez and Wintner (2011).

In our current context, we only give examples of simple feature structures and the output of unification between them. For instance, the following feature structure (FS) has three features: number, gender, and person. This feature structure can represent the agreement features of the Hebrew noun *šwlxn* (pronounced [*shulxan*]) `table'.

number	singular
gender	masculine
person	3

The following FS has 2 features, and it can represent the lexical node of the Hebrew adjective *xwm* [*xum*] `brown.masc'.

$$\begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \end{pmatrix}$$

When unifying the two FSs, the corresponding values of the features are matched, and found to be identical for the features "number" and "gender". Since only the first FS contains the feature "person", the output will contain the same feature-value pair too. The output representing the noun phrase *šwlxn xwm* [*shulxan xum*] `brown table' FS will be:

$$\begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \\ \text{person} & 3 \end{pmatrix} \cup \begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \end{pmatrix} = \begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \\ \text{person} & 3 \end{pmatrix}$$

However, if we change the gender feature of the second FS to *feminine,* so it would match the Hebrew adjective *xwmh* [*xuma*] `brown.fem', the unification would fail, due to contradiction in the value of the feature "gender".

$$\begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \\ \text{person} & 3 \end{pmatrix} \cup \begin{pmatrix} \text{number} & \text{singular} \\ \text{gender} & \text{masculine} \end{pmatrix} = \emptyset$$

This unification-like mechanism is used in Stat-XFER as constraints added to the SCFG rules, in order to validate grammatical characteristics such as agreement, regardless of the adjacency of the related objects in the sentence. As an example, figure 3b below is the Stat-XFER representation of a rule similar to that of figure 3a, mapping an indefinite Hebrew noun phrase into a corresponding English noun phrase, this time with the unification-augmented constraints on the features. The first line is the name of the rule. The second and third lines are optional source-language (SL) and

target-language (TL) examples. The fourth line is the SCFG rule itself, mapping a SL parent constituent of type NP1 and its ordered daughters NP1 and ADJ to a TL parent constituent of type NP1 and its ordered daughters ADJ and NP1. Next is the alignment between daughter nodes of both sides. X marks the source language, Y marks the target language. X1 stands for the first constituent on the source language (here NP1), X2 is the second, etc. The unification-style constraints are then listed. Here the constraints are only on the SL side, but they can also be on the TL side, or between SL and TL. Finally, the parent nodes X0 and Y0 are generated, and their features are set either from one of the daughter constituents (as is the case here), or with explicit unification constraint for each such feature (used in more complex cases).

```
{NP1,2}                              # rule name
;;SL: $MLH ADWMH                     # source language example
;;TL: A RED DRESS                    # target language example
NP1::NP1 [NP1 ADJ] -> [ADJ NP1]      # SCFG rule
(X2::Y1)                             # alignment of SL and TL constituents
(X1::Y2)
((X1 def) = -)                       # unification constraints on SL side
((X1 status) =c absolute)
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
(X0 = X1)                            # propagation of features to the output FSs
(Y0 = Y1)
```

Figure 3b: A Stat-XFER rule that maps a Hebrew indefinite noun phrase to its English counterpart, using unification-augmented constraints.

3.5 Grammar

A grammar consists of a collection of synchronous context free rules, which can be augmented by unification-style feature constraints. These transfer rules specify how

phrase structures in a source-language correspond and transfer to phrase structures in a target language, and the constraints under which these rules should apply.

3.6 Engine

During the process of translation, the input sentence is analyzed by the morphological analyzer, which tokenizes the sentence into morphemes and returns a FS for each morpheme. A transfer engine applies SCFG rules of the transfer grammar to the morphologically analyzed SL input sentence. During rule application, the engine uses the bilingual lexicon, and creates the TL hypotheses. This process operates on both SL and TL simultaneously in a bottom-up fashion. This means that the generation of the output translation is also done bottom-up.

During rule application, the unification constraints are checked, and only valid FSs are generated. After each successful rule application, the TL constituent is added to the lattice, which contains collections of scored word- and phrase-level translations (hypotheses) according to the grammar.

3.7 Decoding

As a final stage, the decoder chooses the most adequate hypothesis according to a statistical score. Several statistical features participate in score calculation, with weights assigned to each of them. In Stat-XFER, these features are the LM score of the target hypothesis, the number of different fragments in the hypothesis, and the ratio between output and input lengths. The weights of these features can be either manually configured or learned using machine learning techniques. The hypothesis score is log-linear with these features and their weights. This means that the score is of the form:

(5) $$P(t|s) = \frac{1}{Z} \cdot \exp[\sum_1^N \alpha_i \cdot h_i(t,s)]$$

where $h_i(t,s)$ is a feature function which outputs a number given the input strings of t and s, and α_i is the weight of the feature in the total score. Finally, $\frac{1}{Z}$ is a normalizing factor in order to get a probability function that sums to 1.

Since the decoder cannot go over every possible hypothesis, a certain number of best hypotheses to examine (called *beam width*) is determined, and further hypotheses are ignored. This is a crucial factor, since in a scenario where there is a very large number of hypotheses (also called *lattice explosion*), correct translations may not be scanned at all, due to computational reasons only.

Since generation is performed bottom-up, the number of hypotheses generated is greater than the number of correct hypotheses, since agreement and word re-ordering may take place only at a later stage. Therefore, earlier generations are often over-generations, which can be ignored only at later stages upon construction of longer constituents. For example, all forms of the TL verb are generated by the generator when processing the SL verb. These TL verb forms are added to the lattice, and only at a later stage during the bottom-up parse process when the subject and the verb participate in the same sentence-level rule, can the agreement features of the verb be verified using unification (a similar approach is detailed for translation of irrational plural nouns in Arabic in section 7.2.2). Such over-generation is a common reason for ungrammatical output of the decoder.

3.8 Stat-XFER as a translation platform

Translation scholars (see, e.g., Lörscher, 1991) have shown that more experienced translators shun away from the wording of the original texts and tend to render the text more freely, keeping equivalence with meaning rather than form. In contrast, novice translators render the text on the go, bound by the original wording and phrasing of the original. We can loosely set the analogy according to which experienced translators follow the interlingua paradigm, whereby a human/machine first reads and understands the meaning of a sentence, and only then translates the meaning into its most adequate representation in the target language

The Stat-XFER approach to translation is different in many ways. First of all, the translation process is done bottom up simultaneously. This means that there is no full parse (syntactic or semantic) of the source sentence before the output sentence or any of its particles are created. Moreover, there is no semantic representation of meaning whatsoever, and the mapping between source and target sentences is done based

solely on morpho-syntactic features, the bilingual lexicon, and a statistical n-gram model of the target language.

The SCFG maps input structures to target structures. There are different types of correspondences and constraints being used in the grammar.

1. The first is purely linguistically motivated. For example, this includes word- and constituent-reordering, which is a linguistic phenomenon often easily mapped between the two languages. Other examples include enforcement of agreement constraints in the source language or in the target language.

2. The second type of correspondence has linguistic reasons, but is actually statistically motivated. If there is a distributional similarity between morpho-syntactic features of the two languages (such as definiteness, plurality, gender, etc.), we can propagate these features from the source structure to the target structure, even if this is not always linguistically licensed. For example, in many cases, a plural noun in Hebrew will be translated into a plural noun in Arabic, but this is not always true. There are plural nouns in Hebrew that are translated into mass nouns (syntactically singular but semantically plural). One such case is
   tpwxim [*tapuxim*] → *tFAhp*
   `apple.pl' (Hebrew) `apples.mass' (Arabic).
   The mapping of the number and definiteness features is done for statistical reasons, since these features are often identical between the two languages.

3. Some constraints are added for computational reasons, in order to decrease the number of hypotheses needed to check. The bottom-up direction of translation enforces us to prune hypotheses during translation, and some decisions should be taken at an early stage to avoid lattice explosion.

Stat-XFER enables the grammar writer to map the source and target constructions, while using external linguistic knowledge. This knowledge, given its availability, can help in solving intricate linguistic issues in translation.

3.8.1 Gender

Syntactic gender is the gender related to a noun in a language, regardless of any real-world attribute of the noun itself. Natural gender, however, is the real-world gender of the object itself. Syntactic gender is idiosyncratic and should not be transferred. For example:

šwlxn [*shulxan*] → *TAwlp*
`table.masc' (Heb) `table.fem' (Arabic).

However, natural gender should usually be transferred, especially when there are both masculine and feminine forms of the noun. For example:

nšia [*nasi*] */ nšiah* [*nesi'a*] → r}ys */ r}ysp*
`president.masc/fem' (Hebrew) `president.masc/fem' (Arabic).

The decision of whether the gender of the source noun should be transferred or not is therefore not only syntactic but also semantic. Statistical direct-translation approaches map input and output strings, and can therefore perform the correct transfer of gender implicitly (though such systems often make mistakes in enforcing agreement since they do not model agreement features directly). However, transfer-based approaches map representations of the input based on lemmas (not on surface forms), and require the mapping of the features explicitly. Currently, with no available resources regarding the natural gender of nouns for Hebrew and Arabic, the lemmas may be correctly translated, but the natural gender is not transferred, in order not to hamper the transfer of different syntactic gender. For example, in order to preserve the correct translation of *šwlxn* `table.masc.' (Hebrew) → *TAwlp*,`table.fem.' (Arabic), one cannot correctly translate *nšiah* `president.fem' (Hebrew) → *r{ysp.*`president.fem' (Arabic). Incorporating the knowledge of which nouns have a natural gender as a feature in the lexicon can help in properly translating nouns with a natural gender.

3.8.2 Mass nouns vs. count nouns

As previously mentioned, morphological number values are often similar between languages in translations of the same sentence, but this need not be true. In some cases, a morphologically-plural count noun would be translated into a morphologically-singular mass noun, despite both nouns being semantically plural. For example, *tpwxim* `apple.pl' (Hebrew) → *tFAhp* `apples.mass' (Arabic). Similar to the proposed solution to maintaining differences in gender, in order to enable different morphological numbers in the two languages, one can incorporate external linguistic knowledge in the lexicon, and use it to allow correct translation while properly enforcing agreement. Unfortunately, the information regarding the countability of nouns required to enable such a difference in the input and output sentences is not available.

3.8.3 Definiteness

The decision of whether to add a definite article before noun varies between languages. In our case, empirical results showed that the choice of definiteness is almost always identical, but there are some cases in which this is not true. We currently propagate the definiteness of NPs from source to target language, since this is statistically true. However, for languages that are not so closely related, this policy may fail. Correctly translating definiteness in these cases may require a more explicit model of definiteness in the target language.

As we shall see, Stat-XFER can be used to map intricate correspondences between two languages. In order to correctly map subtle differences between two languages, external knowledge may be needed, be it lexical, syntactic or semantic. The dearth of such resources forces us to enforce coarse-grained constraints, which hold for most of the cases.

4. Hebrew and Arabic – Similarities and Differences

Modern Hebrew and Modern Standard Arabic, both closely-related Semitic languages, share many orthographic, lexical, morphological, syntactic and semantic similarities, but they are still not mutually comprehensible[1]. Most native Hebrew speakers in Israel do not speak Arabic, and the vast majority of Arabs (outside Israel) do not speak Hebrew. This chapter surveys the main similarities and differences between these two languages and their effect on MT.

4.1 Orthography

4.1.1 Letters and diacritics

While Hebrew and Arabic use different writing systems, they share many orthographic similarities. Their orthographies consist of a system of letters, denoting consonants and long vowels, and diacritics, which denote short vowels. In both languages, the diacritics are typically omitted in contemporary texts, which leads to high morphological ambiguity, and makes text analysis a harder task.[2]

Hebrew, as oppose to Arabic, does not have a single common way of writing words. Since the diacritics are omitted, many words have several surface form alternations, with short vowels written as letters representing long vowels (*ktiv male*), or omitted all together (*ktiv xaser*). Some words have acceptable surface forms with some short vowels written as letters and other vowels omitted. Unfortunately, there is no perfect mapping between these two writing methods, which poses a challenge for processing Hebrew text.

Translating to non-diacriticized Arabic (or Hebrew) has its advantages, since many variant words share the same non-diacriticized form and differ only in diacritics. For example, distinction in gender in second person pronouns is lost in some scenarios in

[1] In certain respects, Arabic Dialects have morpho-syntactic features closer to Hebrew than Modern Standard Arabic, e.g., the absence of nominal case and verbal mood, the behavior of the feminine ending in genitive constructions, the gender-number invariance of the relativizer, and the dominance of SVO order over VSO order. We do not discuss Arabic dialects here.

[2] To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are abgdhwzxTiklmns‘pcqršt. For Arabic we use the transliteration scheme of Buckwalter (2002). Phonetic forms are given between slashes.

both languages: the Hebrew forms [katavta] ‘you (2.sg.m) wrote’ and [katavt] ‘you (2.sg.f) wrote’ collapse into the non-diacriticized form *ktbt*; and the Arabic forms [baytuka] ‘your (2.sg.m) house’ and [baytuki] ‘your (2.sg.f) house’ collapse into the non-diacriticized form *bytk*. Moreover, Arabic case and mood features, absent in Hebrew, are often realized as diacritics only: e.g., the Arabic orthographic word *wld* ‘boy’ can stand for [waladu] (nom. def.), [waladun] (nom. indef.), and [waladın] (gen. indef.), among others. Also, the distinction between the indicative, subjunctive and jussive imperfective forms of most Arabic verbs is lost in some scenarios when the words are non-diacriticized.

4.1.2 Clitics

In both languages, some prepositions (e.g., *b* ‘in, with’, *l* ‘to’), conjunctions (e.g., *w* ‘and’) and the definite article are attached as proclitics to the following word. Attachment of more than one particle can trigger orthographic modifications. For example, Hebrew *b+h+kth* ‘in+the+classroom’ is written *bkth*; and Arabic *l+Al+qlm* ‘for the pen’ is written *llqlm*. Arabic attaches pronominal direct objects as post-verbal clitics, a construction that, while grammatical, is rarely used in contemporary Hebrew. Hebrew uses the definite direct object marker *at* instead.

1. raiti awtm
   ra'iti aotam
   see.1sg.past def.acc+they.acc
   ‘I saw them’ (Hebrew)

2. rAythm
   rAyt +hm
   see.1sg.past they.acc
   ‘I saw them’ (Arabic)

4.2 Word formation

As in other Semitic languages, most nouns and verbs are built from a lexical root, a morpheme consisting of consonants only, which generally denotes a vague semantic meaning, and from templates that add vowels (and, possibly, also consonants) to the root, yielding a lexeme. Many roots are shared between Hebrew and Arabic. For example, the root *k.t.b* 'write' has the same basic meaning in both languages, but it is used in different templates and yields different lexemes. The past tense, 1st person plural form of the verb 'write' is *ktbnw* in Hebrew, and *ktbnA* in Arabic; the noun 'letter (message)' is derived from the same root, and is written *mktb* (pronounced [*mixtav*]) in Hebrew, and *mktwb* [*maktwb*] in Arabic. However, Hebrew also has *mkwtb* [*mexutav*] 'addresee' which is derived from the same root, which does not exist in Arabic, whereas Arabic has *ktAb* [*kitab*] 'book', which does not have a same-root counterpart in Hebrew.

4.3 Inflectional morphology

4.3.1 Nominal morphology

4.3.1.1 Functional gender vs. formal gender

Nouns and adjectives inflect for number, gender and definiteness. In addition, both languages share the difference between the *formal gender* of nouns, which is the gender according to the surface form (expressed as suffixes), and the *functional gender*, which is the gender that is used in agreement. Cases in which the formal gender differs from the functional gender are called irregulars, although there are many such irregular nouns in both languages, especially in the plural form.

3.a AmtHan → AmtHan+At
test.sg.masc test pl.fem
`test' `tests' (Arabic)

3.b. šwlxn → šwlxn+wt
shulxan shulxanot
table.sg.masc table +pl.fem
`table' `tables' (Hebrew)

4.3.1.2 Number

Arabic nominals have three values for the number feature (singular, plural and dual), whereas the dual form only exists in Hebrew in a few frozen cases and is not productive. Furthermore, Arabic has an irregular way of producing the plural form of nouns (called the 'broken plural'), in which no plural suffix is attached to the singular form, and the root is incorporated into one of many morphological templates which are typical of the Arabic broken plural (4c). Even though it is regarded as an irregular form, the plural form of about half of the nouns in Arabic is generated this way. On the other hand, Hebrew plural forms are usually derived from their singular counterparts by adding a plural suffix (4a).

4.a ild → ild+im
yeled yeladim
boy.sg boy+pl.masc
'boy' `boys' (Hebrew)

4.b lAEb → lAEb+wn
actor.sg actor+pl.masc
`actor' `actors' (Arabic, regular plural)

4.c wld → AwlAd
boy.sg boy.pl
`boy' `boys' (Arabic, broken plural)

4.3.1.3 Case

Another important difference between the two languages is that Arabic encodes case on nouns, whereas Hebrew does not. The difference among different Arabic cases is usually expressed only by different vowels, but it is morphologically overt even in non-diacriticized forms in some cases, like in the following example.

5.a Al+klb *hb fy Al+šArE
the+dog.nom walk.past.3ms in the+street
`The dog walked in the street' (Arabic)

5.b Al+wld AH* klbA
the+boy.nom hold.past.3ms dog.acc.indef
`The held a dog' (Arabic)

4.3.1.4 Status

Nominals typically come in three varieties (called *states*): *absolute*, *definite* and *construct* state, which is used in genitive constructions (see Section 4.4). Feminine nouns in the construct state behave differently: In Hebrew such forms trigger a change of the feminine ending *-h* to *-t*. In Arabic, the feminine ending in construct states combines the duality of *-h* and *-t*, since it is written using *-h*, but pronounced as *-t*. Moreover, in both Arabic and Hebrew, the feminine suffix *-h* changes orthographically into *-t* before a possessive pronominal enclitic. For example, in Hebrew the feminine noun *xtwlh* 'cat' changes in this construction into *xtwlt rxwb* 'street cat'; In Arabic, *qTh* 'cat' in construct state is *qTh $ArE* 'street cat', which is written with *-h* but pronounced with a *-t*. The final feminine suffix *-h* changes into *-t* in the possessive form: *qTtnA* 'our cat' (Arabic), *xtwltnw* `our cat' (Hebrew). Construct state inflection in Arabic and Hebrew is similar in other cases, e.g., the regular masculine plural suffix +im (Hebrew) and +wn/+yn (Arabic) is shortened to +i in Hebrew and +w/+y in Arabic.

6.a iwm → imim
yom yamim
day.sg day.pl
`day' `days' (Hebrew, absolute state)

6.b iwm hwldt → imi hwldt
yom huledet yemei huledet
day.sg birth.sg day.pl.constr birth.sg
`birthday' `birthdays' (Hebrew, construct state)

6.c mmvl → mmvlwn

representative.sg representative.pl

`representative' `representatives' (Arabic, absolute state)

6.d mmvl Al$Arkp → mmvlw Al$Arkp

representative.sg company.def representative.pl company.def

`the company representative' `the company representatives' (Arabic, construct state)

4.3.1.5 Pronouns

Many similar pronouns are common to both languages, and pronouns inflect for the same features (number, gender, person and case). This makes translation of pronouns easier. Both nouns and prepositions can combine with cliticized pronominal suffixes that encode number, gender and person (of the possessor or the object of the preposition), e.g., *lnw* 'to us (Hebrew)', *lnA* 'to us' (Arabic).

4.3.2 Verbal morphology

4.3.2.1 Forms

Verbs inflect for number, gender, person and tense, and the two languages share a complex and similar verb structure and inflection system. The two languages share the same verbal forms:

1. The perfective form is used for the past tense in Arabic and Hebrew
2. The imperfective is used for the future tense in Hebrew but is used for a variety of tenses in Arabic (past, present and future) in coordination with various moods and particles.
3. The imperative
4. The active and passive participles are used for present tense in Hebrew and to a lesser extent as a deverbal in Arabic.

The ambiguity of the Arabic imperfective form is a challenge for translation since it can correspond to multiple Hebrew forms: the negated forms of the Hebrew

ktb/kwtb/iktwb ‘he wrote/writes/will-write’ translate to Arabic *lm/lA/ln yktb*, all using the same verb with different moods and particles combining tense and negation (in the case of *lm* and *ln*).

Another difference between the two languages in the imperfective are the two Arabic morphemes denoting future tense: The proclitic *s+* and the separated morpheme *swf* are used in Arabic before the imperfective verb to mark future tense, distinguishing it from the present/future tense interpretation of the verb without these morphemes. For example, *syktb* `(he) will write’. There is no parallel Hebrew construction.

4.3.2.2 Templates

As mentioned in section 4.2, both languages share a system of verbal templates, in which three or four consonants yield a root, which is incorporated into the template to comprise a verb. Every such template has a different and unique inflection paradigm for the five forms mentioned above (perfective, imperfective, imperative, active participle and passive participle). Hebrew has seven different verbal templates, while Arabic has nine common ones (but other rare templates exist). The templates represent, to a certain extent, semantic attributes of the verb, such as the unmarked simple verb, the intensified verb, passivization, causativization, the decausative, the reciprocal, etc. Unfortunately, there is no direct mapping between the different templates in both languages. Moreover, there are many exceptions to the generalizations regarding template-meaning mapping in each language (Ornan 2003).

4.3.2.3 Passivization

Passivization is implemented differently in the two languages. Hebrew predominantly employs a morphological mechanism whereby an active verbal pattern has a passive counterpart in another verbal template. This is highly productive for two template pairs (*pi‘el–pu‘al* and *hif‘il–huf‘al*), less so for the third (*pa‘al–nif‘al*).[3]

Arabic utilizes a different mechanism of vowel change, which is productive for almost all verbal patterns. In this mechanism, the pattern of vowels in the verb is

[3] The seventh template in Hebrew (hitpa'el) is used for reflexive, decausative and reciprocal semantic meanings, and has no passive counterpart.

changed in a consistent manner for each verbal template for each tense. Since the change is in vowels which are omitted in the standard orthography and are not morphologically overt, identifying passive voice can only be done using the context surrounding the verb.

4.3.2.4 Shared ambiguity

In both Hebrew and Arabic, the second person singular masculine and third person singular feminine forms are homonymous across the verbal paradigm in the imperfective/future tense. For example, *tktwb* ‘you.sg.m/she will write’ (Hebrew), *tktb* ‘you.sg.m write/she writes’ (Arabic). This is a clear case of morphological ambiguity that does not have to be resolved in translation.

4.4 Syntax

4.4.1 Word order

The dominant word order is SVO in Hebrew, VSO in Arabic (although other orders are possible), but there are some syntactic constraints on this default order. In Arabic, an embedded clause after the subordinating conjunction *An* must start with a noun (the subject if it is definite, or an expletive pronoun if the subject is indefinite). In addition, the subject of the clause should be in accusative case. Hebrew has no parallel construction. On the other hand, when a Hebrew sentence begins with an adverbial, the default order is VSO.

4.4.2 Agreement

Both Arabic and Hebrew have a complex agreement system, involving features such as person, number, gender, and definiteness. In both languages agreement constraints hold between the following POS pairs:

4.4.2.1 N-Adj

When an adjective modifies a noun, they should agree on number, gender and definiteness. NP-internal word order is identical in both languages.

7.a h+ild h+gbwh
ha+yeled ha+gavoha
the+boy.sg.m the+tall.sg.m
`The tall boy' (Hebrew)

7.b Al+wld Al+Twyl
the+boy.sg.m the+tall.sg.m
`The tall boy' (Arabic)

A peculiarity of Arabic is that the agreement features of plural, irrational (non-human) nouns are always singular feminine, regardless of the gender of the singular noun, and ignoring the semantic plurality of the noun. Every reference to that noun in the sentence must agree with these features:

8.a Al+qlm Al+jmyl
the+pen.m.sg the+pretty.m.sg
‘The pretty pen’ (Arabic)

8.b Al+AqlAm Al+jmylp
the+pen.m.pl the+pretty.f.sg
‘The pretty pens’ (Arabic)

4.4.2.2 Subject–verb

In both languages the verb and the subject NP agree on person, number and gender. However, in Arabic VSO sentences the verb is always in singular:

9.a ktb Al+AwlAd
write-past.sg.m boy-pl.m.def
‘The boys wrote’ (Arabic)

9.b h+ildim ktbw
ha+yeladim katvu
boy-pl.m.def write-past.sg.m
‘The boys wrote’ (Hebrew)

4.4.2.3 Verbless predicates

Both languages have a common construction of verbless sentences, where the predicate is either a PP, another NP or an adjectival phrase. In both latter cases, the subject and the predicate must agree in number and gender, but the subject must be definite and the predicate indefinite:

10. Al+wld Twyl
boy.m.sg.def tall.m.sg.indef
‘the boy is tall (Arabic)

4.4.2.4 Genitive constructions

In both languages a noun–noun construction (called *smikhut* in Hebrew, *idafa* in Arabic) is used to express genitive relations. The head of the structure is the first noun, which determines the number and gender agreement features. The definiteness of this structure is marked on the second noun only.

11.a iwm h+hwldt
yom ha+huledet
day.m.sg.indef the+birth.f.sg
‘The birthday’ (Hebrew)

In Hebrew, but not in Arabic, such relations can also be expressed in a different construction, using the possessive preposition *šl* ‘of’.

11.b h+spr šl h+ild
ha+sefer shel ha+yeled
the+book.sg of the+boy.sg
‘The boy's book’ (Hebrew)

Hebrew exhibits yet another construction of double genitives, which does not exist in Arabic. In this construction, the antecedent noun is followed both by a cliticized possessive pronoun and by a PP headed by *šl*.

11.c spr +w šl h+ild
sifro shel ha+yeled
book.sg+his of the+boy.sg
‘The boy's book’ (Hebrew)

4.4.2.5 Quant–N

Subtle agreement constraints hold between quantifiers (e.g., numerals) and the nouns they modify. These constraints differ across the two languages.

4.4.3 Pro-drop

In both languages, a subject pronoun can be omitted if the verb is in past, future or imperative forms. The agreement features of the subject can be deduced from the morphological form of the verb. This may facilitate translation in some cases: target pronouns do not have to be explicitly generated when they are missing in the source language.

12. AjtmEtm b+Al+DAbT
meet.past.2mp in+the+officer
`You met the officer' (Arabic)

4.4.4 Relative clauses

In Arabic, the relativizer carries gender and number features, and has to agree with the antecedent noun modified by the relative clause. In the following sentence, the relativizer and the encliticized pronoun agree with the antecedent irrational plural noun, and therefore are feminine singular:

13.a Al+AqlAm Alty A$try+hA Al+wld
pen-m.pl.def REL.f.sg buy-past.3.m.sg+she-acc boy-m.sg.def
'The pens which the boy bought' (Arabic)

Such relative clauses modify only definite nouns, as in example 13.a. Relative clauses that modify indefinite nouns have no relativizer, as in example 13.b.

13.b. rAyt wldA qrA ktAbA
see.1st.sg.past boy.sg.m.indef read.3rd.sg.past book.sg.indef
'I saw a boy [who] read a book' (Arabic)

In Hebrew relative clauses usually use the lexical relativizer *š*, which carries no agreement features.

13.c. raiti ild š qra spr
ra'iti yeled she kara sefer
see.1st.sg.past boy.sg.m.indef REL read.3rd.sg.past book.sg.indef
'I saw a boy who read a book' (Hebrew)

Arabic extensively uses resumptive pronouns in relative clauses for both indirect and direct object pronouns (14a). In the common Hebrew clause that starts with a lexical relativizer this is arguably ungrammatical for direct objects pronouns (14c) and not used in texts.

14.a Al+wld Al*y rAyt +h
boy.def REL.m.sg see.past.1s+he.acc
`The boy I saw' (Arabic)

14.b h+ild š raiti
ha+yeled she ra'iti
boy.def REL see.past.1s
`The boy I saw' (Hebrew)

14.c ?? h+ild š raiti awtw
ha+yeled she ra'iti oto
boy.def REL see.past.1s he.acc
`The boy I saw' (Hebrew)

However, there is a construction in Hebrew where a resumptive pronoun acts as a relativizer. This exists for both direct and indirect object pronouns.

14.d h+ild awtw raiti
ha+yeled oto ra'iti
the+boy he.acc see.past.1s
`The boy whom I saw' (Hebrew)

14.e h+kdwr b+w šixqti
ha+kadur bo sixakti
the+ball in+he.gen play.past.1s
`The ball which I played with' (Hebrew)

Arabic has no parallel construction.

Hebrew also has a construction in which the relativizer is the definite article *h*+, which can be used in this function only if the embedded verb is in the present. A similar phenomenon in Arabic uses the definite article with the active participle deverbal form.

15.a h+mkwnit h+ xwnh

ha+mexonit ha xona

the.car REL park.active_ptcp

`The parking car' (Hebrew)

15.b Al+syArp Al+ mtwqfp

the.car REL park.active_ptcp

`The parking car' (Arabic)

5. Challenges

The similar characteristics of Arabic and Hebrew can indeed be beneficial for MT, but the differences listed above pose some intricate challenges. In this chapter, some of those challenges are listed. In the next chapter, possible solutions to these issues are suggested.

5.1 Orthographic challenges

As mentioned in section 4.1.1, Hebrew does not have a single common convention of writing words, and the transfer between *ktiv male* and *ktiv xaser* is not consistent. The Hebrew side of our bilingual dictionary is written using *ktiv xaser*, while most texts are written using some sort of *ktiv male*. For this reason, the ambiguity in analysis of Hebrew rises, and the matching of text words to bilingual entries is problematic: words that appear in our bilingual lexicon in *ktiv xaser* may not be matched to an analysis of a word which is written in *ktiv male*.

5.2 Lexical challenges

As in other language pairs, Hebrew and Arabic verbs have different subcategorization frames for corresponding verbs. Some Hebrew verbs require a specific preposition before the (indirect) object while in Arabic the object is direct, and vice versa.

16.a nkx b+ h+pgišh
Naxax b+ a+pgiša
attend.3sg.m.past in+ meeting.def
'He attended the meeting' (Hebrew)

16.b HDr Al+jlsp
attend.3sg.m.past meeting.def
'He attended the meeting' (Arabic)

This phenomenon is of course not special to Hebrew-Arabic. However, combined with differences in word order between the two languages, its effect is enhanced. While during the process of decoding the language model may help to correctly choose the preposition in the Arabic output sentence based on the local context, this is less likely in sentences with long distance V–O dependencies, since the subject and other adjuncts may intervene between the verb and its preposition.

17. AErb r}ys Al+Hkwmp ywm Al+ArbEA' fy jlsp
express.3sg.m.past leader government.def day Wednesday in meeting
Al+Hkwmp Al+AsbwEyp En Aml +h …
government.def weekly.def upon hope he.poss
'The prime minister expressed Wednesday during the government weekly meeting his hope …' (Arabic)

This example demonstrates the potential distance between the verb *AErb* 'express' and its required preposition *En*, which are separated by the subject NP and other temporal and locative adjuncts. This distance hampers the ability of a statistical, n-gram-based language model to correctly select the preposition.

Another lexical challenge stems from the fact that existing Arabic lexical resources (Buckwalter, 2004; Habash, 2004) do not encode information on functional gender and rationality of nouns, which is crucial for enforcing N-Adj agreement. The implication is that in order to generate Arabic, one must over-generate both masculine and feminine forms, delegating the choice to the language model, which chooses poorly in long-distance dependencies.

5.3 Morphological challenges

Translating between two morphologically rich languages poses challenges in morphological analysis, transfer and generation. The complex morphology induces an inherent data sparsity problem, and the limitation imposed by the dearth of available parallel corpora is magnified (Habash and Sadat, 2006).

As a specific example, consider passive verbs. Since passivization in Arabic is expressed as vowel changes, it is usually not morphologically overt and is harder to identify. The passive form in Hebrew is not fully productive (especially in *Pa'al-Nif'al* template pair), and is not always predictable. Therefore, both the identification and the generation of the passive voice pose a challenge in translating into the correct form of the verb.

5.4 Syntactic challenges

5.4.1 Word order

Arabic word order is relatively free, as in Hebrew. This means that there are many possible correspondences between Hebrew and Arabic word orders. Since the dominant word order in Arabic is VSO, the verb and its object are not necessarily consecutive. As a result, the variability of possible sentence structures has to be accounted for on the sentence level, rather than on levels such as VP.

Generating the correct word order in an embedded clause that starts with *An* (see Section 4.4.4) is a complex issue. It requires generation of several different structures at the embedded sentence level, forcing subtle order constraints according to the embedded sentence structure, and afterwards (when the relative clause is combined with the relativizer) validating that this was indeed inside an embedded clause.

5.4.2 Mapping unique constructions

A major challenge stems from syntactic constructions and word formations that have no counterpart in the other language. When these constructions appear in the source language, they need to be mapped into a matching parallel construction in the target language. In the other direction, it may be necessary to generate such constructions in the target language which do not exist in the source language. This can be even harder, since some morpho-syntactic features (e.g., case) need to be properly determined and generated in the target language without knowing their value in the source language.

For example, the Hebrew *šl* genitive (example 11b) and double genitive (11c) constructions do not directly correspond to an Arabic construction.

When translating into Arabic, the Hebrew cliticized possessive pronoun *+w* and the genitive prepositions *šl* must be omitted, and the corresponding Arabic *idafa* structure has to be generated (18) with the proper case assignment.

18. ktAb Alwld
    book.sg the+boy.sg
    'The boy's book' (Arabic, idafa)

Another Hebrew particle that has no parallel in Arabic is the accusative definite marker *at*. When translating into Hebrew, the correct case of the Arabic NP needs to be determined, and the correct inflection of *at* needs to be generated explicitly: when the Hebrew accusative marker *at* is followed by a pronoun, they are merged together creating a single morpheme. For example, *awtw* `him', *awtm* `them'.

From our empirical work, it seems that more constructions in Arabic do not exist in Hebrew than the other way round. The example of Arabic case was previously mentioned as a morpho-syntactic feature that is usually not morphologically marked in Hebrew and needs to be determined for correct Arabic generation. Another example is the verbal mood. In relative clauses and in verbal constructions following certain prepositions, the Arabic imperfective verb inflects differently and takes the subjunctive or jussive moods, which do not exist in Hebrew. These constructions need to be identified and dealt with correctly during the generation process. Another example is the Arabic *Dmyr Al$An* construction – the expletive pronoun. Relative clauses following a closed set of prepositions need to start with a noun. If the subject of the clause is definite, it will start the clause. Otherwise, an expletive pronoun *+h* opens that clause.

19. qAl Al+wzyr An **+h** lA ymkn Ayqaf Al+bAxrp
    say.past.3ms the+minister that expl. not possible stop the.ship
    `The minister that it is not possible to stop the ship' (Arabic)

There are many other such constructions that exist only in one language. Mapping, analysis and generation of these constructions poses a challenge to any rule-based MT system between two different languages.

5.4.3 Enforcing agreement

As we have shown in section 4.4, Arabic poses many syntactic challenges in correctly enforcing agreement. For example, subject–predicate agreement in verbless sentences whose predicate is an adjectival phrase requires identification of the heads of the subject NP and the (potentially distant) indefinite adjectival predicate, and forcing agreement between them:

20. Al+wld Al*y rAyt +h fy Al+mTAr Al+kbyr Twyl
boy.sg.m.def REL.sg.m see.1.sg.past he.acc in airport.m.def big.m.def tall.m.indef
'The boy I saw at the big airport is tall' (Arabic)

In the case of subject–verb agreement on number, when the Arabic form of the verb is generated, it is unknown whether the verb will be placed before or after the subject. This poses a challenge for generating the correct form of the verb.
A more complex issue is the plural form of irrational nouns in Arabic. As demonstrated in (13a), any reference to such a noun must use singular feminine agreement features. This requires information about the irrationality of the plural noun, the particles that need to agree with it, and enforcement of long distance agreement.

5.5 Computational challenges

Every MT system handles the problem of potential lattice explosion. The problem is enhanced in translating from and to morphologically rich languages, such as ours. The lack of a morphological disambiguator during analysis enhances this effect. This issue is especially true in the case of our systems, which process both the source and the target languages bottom-up simultaneously, in order to prune target hypotheses during parsing. Some syntactic choices are determined only at relatively late stages, resulting in huge hypothesis spaces earlier. For every verb, the Arabic morphological generator

returns 109 possible forms (excluding possible clitics). This is the number of possible results out of the cartesian product of several many-valued morpho-syntactic features: person, gender, number, aspect (perfective, imperfective and imperative), voice (passive or active), and mood (indicative, subjunctive or jussive). For every noun, 72 forms are returned (excluding possible clitics), as a result of the various values of the features gender, number, case, possessiveness and definiteness. The implication is that the number of wrong hypotheses in the lattice is very large. Since only a fraction of these hypotheses can be examined and further inspected, many hypotheses need to be ignored, based on a limited scoring function. Correct translations are often filtered out this way, which yields ungrammatical output translations.

6. Possible solutions

As the standard paradigm of statistical MT is not applicable to Hebrew-to-Arabic MT, due to the dearth of available parallel corpora, two alternatives present themselves. One is translating using a third language (most naturally, English) as a pivot (Muraki, 1987; Wu and Wang, 2007); the other is relying on linguistically-motivated transfer rules, augmented by deep linguistic processing of both the source and the target languages[4]. Both approaches are considered below.

6.1 Using English as pivot

The dominant Hebrew-to-Arabic MT system is Google's.[5] Google has been known to use 'bridge' languages in translation (Kumar et al., 2007). We provide evidence that Google's Hebrew-to-Arabic MT uses English as a pivot, and demonstrate the shortcomings of this approach.[6] As a first test, we use the number- and gender-ambiguity of second-person pronouns in English (you). Since Hebrew and Arabic use separate forms for these pronouns, direct translation is not expected to be ambiguous; however, Google produces the following wrong translations in such cases (Hebrew on the left, Arabic on the right of the arrows):

21.a atm / atn → Ant
you.pl.m / you.pl.f you.sg.m/f
`You' (Hebrew to Arabic)

21.b qlt l+km → amrti lk
say.1sg.past to+you.2.pl.m-dat say.1sg.past to+you.2.sg.m/f-gen
`I told you' (Arabic to Hebrew)

[4] A third approach is to use comparable corpora (Munteanu and Marcu, 2005); but with no parallel data whatsoever, this is unlikely to succeed.
[5] http://www.google.com/language_tools, accessed May 5th, 2010
[6] Another Hebrew-to-Arabic MT system, http://www.microsofttranslator.com/, also uses English as a pivot language, and shows similar characteristics.

21.c klb +km → Alklb

dog.sg+poss.2.pl.m dog.sg.def

'Your dog' → 'the dog' (Hebrew to Arabic)

The second test uses the fact that plural nouns in English are unspecified for gender, whereas in Hebrew and Arabic they are. Here, gender is lost in translation of plurality, and the decoder chooses the most common option according to the language model.

22. mwrim / mwrwt → mElmyn

teachers.m / teachers.f teachers.m

In the third test, we translate words which are lexically ambiguous in English but not in Hebrew or Arabic, such as *table, bank*, and *manual*.

23.a Tblh → TAwlp

table (data) table (furniture)

23.b bnq → sAHl

bank (financial) bank (shore)

23.c idni → ktyb

manual (by-hand) manual (booklet)

The implication of using a morphologically-poor language as a pivot in translating between two morphologically-rich languages is that much data is lost, and the output tends to be either wrong or ungrammatical. The following example summarizes the problems.

24. mwrwt ipwt aklw →

teacher.pl.f.indef pretty.pl.f.indef eat.3.pl.past

Aklt AlmElmyn jmylp

eat.3.sg.f.past teacher.pl.m.acc/gen.def pretty.sg.f.indef

'Pretty teachers ate' → 'Teachers ate pretty' (Hebrew to Arabic)

The following issues can be observed:

1. Gender mismatch (feminine *mwrwt* vs. masculine *AlmElmyn*). The reason is that English nouns are unspecified for gender.
2. Number mismatch (plural *ipwt* and singular *jmylp*). This results in the wrong translation and a disfluency in the target sentence. The reason is that English adjectives are unspecified for number.
3. Definiteness mismatch (The Hebrew noun is indefinite while in Arabic the noun is definite and the adjective is not).
4. Case mismatch: Hebrew is unspecified, Arabic is accusative/genitive (as opposed to the correct nominative case).
5. Verb conjugation error: the verb that precedes the plural subject *AlmElmyn* is in feminine singular form, although the subject is rational plural masculine.

The errors in properly transferring the correct morphological features and enforcing the syntactic constraints rendered the output sentence incoherent.

6.2 Transfer-based translation

As an alternative to using English as a pivot language, we chose a knowledge-based approach. A linguistically-aware transfer approach has several advantages in this case. Source language morphological analysis provides a tokenization and analysis of the input sentence into morphemes with their morpho-syntactic features. Then, transfer rules and a transfer lexicon map source words and (linguistic) phrases into the target language, bridging over syntactic differences across the languages. Finally, a target-language morphological generator creates inflected morphemes from the yield of the target tree fragments; a subsequent detokenization step then recreates the correct orthographic forms. We use the Stat-XFER framework (Lavie, 2008), which uses a declarative formalism for symbolic transfer grammars.

7. Transfer-based SMT systems for Hebrew and Arabic

Using the Stat-XFER framework, we successfully implemented a transfer-based MT system for each direction, solving many of the problematic issues raised above, focusing on gapping morphological differences and enforcing agreement.

7.1 Resources

For translation from Hebrew to Arabic, we use a morphological analyzer (Itai and Wintner, 2008) for the Hebrew source, with no morphological disambiguation module.[7] This causes many wrong analyses to be processed and dramatically increases the size of the hypothesis lattice. For generation we use Habash (2004) which requires proper specification of morpho-syntactic features in order to generate the correct inflected form. Clitics are generated separately and are then attached as a post-process (El Kholy and Habash, 2010). In the Arabic-to-Hebrew direction we use Habash (2004), this time as a morphological analyzer and disambiguator. This helps us reduce the number of hypotheses in the lattice. For generation we use the reverse direction of Itai and Wintner (2008) as a generator, which yields better gender-inflected outputs than its Arabic counterpart. Due to the morphological disambiguator in Arabic and the generator in Hebrew, translation in this direction currently performs better. The grammars for both systems were manually crafted, based on prototypes done by Lavie (2004b). The grammars comprise of about 50 rules, 20 of which are dedicated to enforcing correct NP structure, while the other rules mainly deal with sentence level agreement and reordering.

7.2 Transfer rules

The systems correctly generate and decode both Arabic and Hebrew NP-internal structure, verbs with encliticized object pronouns, agreement between subject and adjectival-predicate, and subject–verb agreement (in number, gender and person). We

[7] Such a module is currently under development. Experiments with available POS taggers resulted in poorer performance.

also correctly implemented translation of structures that do not exist in the target language, such as the Hebrew definite accusative marker *at*, the genitive *šl* and double genitive constructions, and the Arabic dual number and future markers *swf* and *s+*. We implemented rules that enforce agreement on rationality and gender between nouns and adjectives, and relate verbs to their subcategorized prepositions; but the large-scale lexical resources needed to fully solve some of these problems are still missing. Most of these rules are applicable for translation in both directions, while some are specific to a certain direction: for example, enforcing agreement for irrational plural nouns is relevant only when translating into Arabic.

Following is a discussion about the solutions implemented for some of the challenges listed in the previous section.

7.2.1 Subject–predicate agreement

In local contexts, this is relatively easy, since a simple rule can use unification constraints to force agreement on all features. When the subject and the predicate (whether verbal, adjectival or nominal) are distant, the agreement features of the head of the subject NP are propagated up the NP, and agreement is checked at the sentence level against the features of the predicate. The main task here is to correctly build the NP and all of its components, including relative clauses. After this process, which is implemented by around 20 different rules, is over, enforcing agreement with the predicate is simple, in a single sentence-level rule. Figures (4a), (4b) below list such a rule, enforcing agreement between subject and adjectival predicate for Arabic rational nouns.

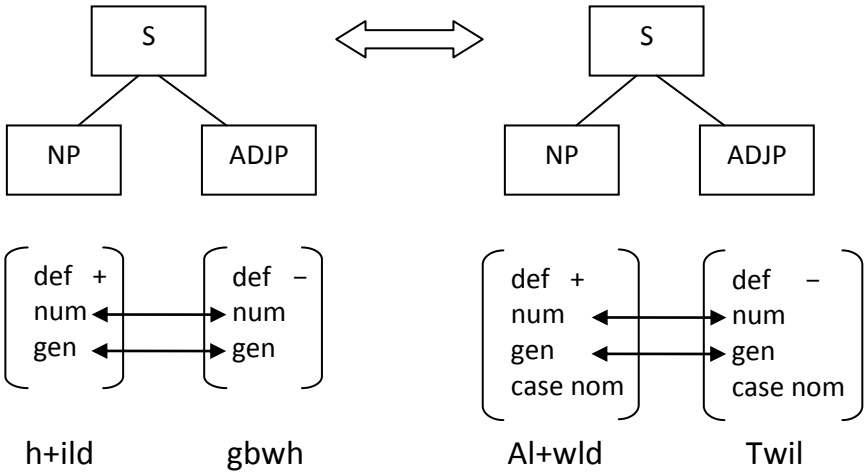

Figure 4a: A tree representation of the adjectival predicate rule

```
{S_NP_ADJ,1}                    # rule name
;;SL: H ILD GDWL               # source example
;;TL: Alwld kbyr               # target example
S::S [NP ADJP] -> [NP ADJP]    # morpheme POS mapping
(X1::Y1)                       # morpheme alignment
(X2::Y2)
((X1 def) = +)                 # Hebrew side agreement
((X2 def) = -)
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
((Y1 rational) = +)            # Arabic side agreement
((Y1 def) = +)
((Y2 def) = -)
((Y1 num) = (Y2 num))
((Y1 gen) = (Y2 gen))
((Y1 case) = nominative)
((Y2 case) = nominative)
```

Figure 4b: Enforcing agreement between subject and adjectival predicate (for rational Arabic nouns), written using Stat-XFER formalism.

7.2.3 NP internal structure

NP is a complex structure in both languages. About 20 rules are responsible for creating this structure, since there are many different NP constructions: Nouns can be either definite or indefinite, in absolute or construct state, proper or common, pronominal or nominal, modified by adjectives or another noun (*smikhut*), and can participate in more complex NP constructions such as genitive constructions or NPs with relative clauses. There are subtle dependencies between these morpho-syntactic features, which makes this task complex. Below are figures (5a) and (5b), which give an example to one of the rules that create NP structure, specifically the *smikhut/idafa* genitive construction that is typical of Semitic languages. This rule reflects one of the structural similarities between Hebrew and Arabic, being totally symmetrical.

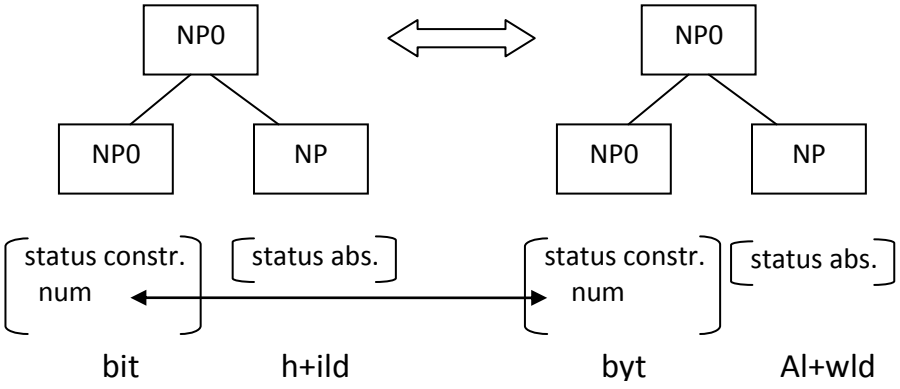

Figure 5a: A tree representation of the s*mikhut/idafa* genitive construction

```
{NP0_SMIKHUT,1}
;;SL: byt (Al+) wld
;;TL: BIT (H+) ILD
NP0::NP0 [NP0 NP] -> [NP0 NP]
(X1::Y1)
(X2::Y2)
((X1 status) = construct)
((X2 status) = absolute)
((Y1 status) = construct)
((Y2 status) = absolute)
((X1 num) = (Y1 num))
```

Figure 5b: A simplified version of the rule that maps *smikhut/idafa* constructions

7.2.4 Genitive constructions

As previously mentioned, the Hebrew genitive using *šl* and double genitive constructions do not exist in Arabic. When translating from Hebrew, these constructions need to be transferred into the parallel Arabic construction using enclitic pronoun (as shown in figures (6a) and (6b)), or into the *idafa* construction. When translating from Arabic, these constructions need to be explicitly generated. In this rule, the *šl* Hebrew genitive particle is omitted, the pronouns are matched for identical morphological features, and the Arabic noun is verified for indefiniteness and possessiveness.

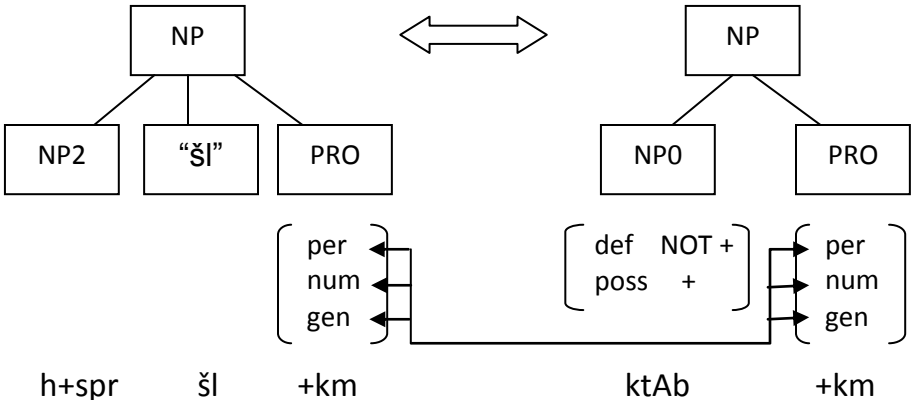

Figure 6a: A tree representation of rule that maps Hebrew genitive constructions using *šl* with a pronoun into an Arabic construction using enclitic pronoun

```
{NP_POSS,1}                  # rulename
;;SL: H SPR $LKM            # source example
;;TL: ktAb +km              # target example
NP::NP [NP2 PREP PRO] -> [NP2 PRO]      # morpheme POS mapping
(X1::Y1)                     # morpheme alignment
(X3::Y2)
((X2 lex) = $L)              # lexical constraint on SL
((Y1 poss) = +)              # syntactic constraint on TL
((Y1 def) = (*NOT* +))
((Y2 per) = (X3 per))        # syntactic constraints on SL-TL
((Y2 num) = (X3 num))
((Y2 gen) = (X3 gen))
```

Figure 6b: A rule that maps Hebrew genitive constructions using *šl* followed by a pronoun into the Arabic *idafa* construction

7.2.2 Irrational plural noun agreement

Recall from examples (8a) and (8b) that irrational plural nouns in Arabic are treated as if they are feminine singular nouns in terms of agreement. Enforcing such agreement is difficult since the morphological features in the FS as returned from the morphological analyzer do not reflect the agreement features.

The naïve solution is to lexically determine the rationality of each noun, and let two different rules generate the verb in the correct form according to the subject's rationality (given that the subject is plural). However, information on rationality is currently not available to us, since it is not morphologically overt and available Arabic morphological analyzers currently do not provide this information. Another solution is to generate both the feminine singular form and the plural form with the original gender of the singular form, and let the language model decide. This may solve the problem in local contexts,

but as was shown in example (20), the phenomenon extends to long-distance dependencies.

Our solution is to combine the two approaches. Two hypotheses are generated, one for the rational form and one for the irrational form. Using the rules, we account for complex NPs with relative clauses, and agreement is enforced among all relevant references to the antecedent noun. By propagating the agreement features up to higher levels of the tree, we guarantee that the predicate agrees with the subject NP, whether it is a regular rational plural or an irregular irrational plural. The two limitations of this solution are the limited syntactic coverage of the rules, and the limited number of hypotheses that can be examined. For long and complex NPs this solution is suboptimal, since the derivation of the full NP does not always succeed. For simpler NPs with short or no relative clause at all, this solution works well.

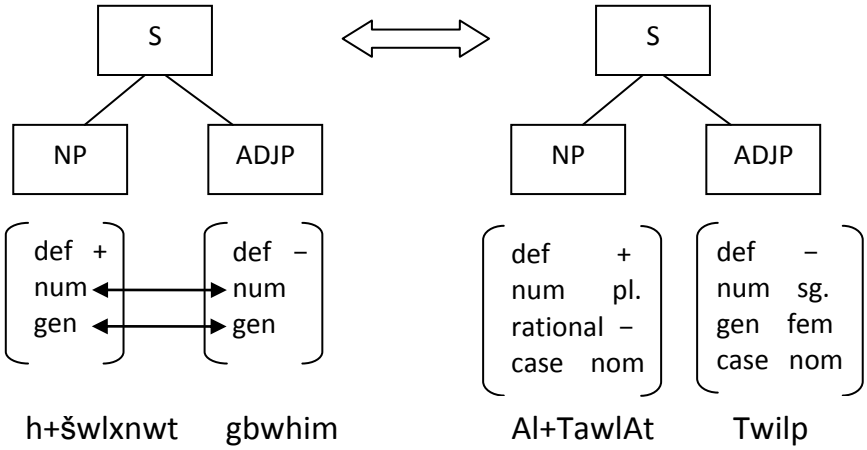

Figure 7a: A tree representation of the rule mapping sentences with adjectival predicates for irrational plural nouns

```
{S_NP_ADJ_IRRAT,1}
;;SL: H$WLXNWT GDWLIM
;;TL: AlTAwlAt kbyrp
S::S [NP ADJP] -> [NP ADJP]
(X1::Y1)
(X2::Y2)
((X1 def) = +)            # Hebrew side agreement
((X2 def) = -)
((X1 num) = (X2 num))
((X1 gen) = (X2 gen))
((Y1 rational) = -)       # Arabic side agreement
((Y1 def) = +)
((Y2 def) = -)
((Y1 num) = plural)
((Y2 num) = singular)
((Y2 gen) = feminine)
((Y1 case) = nominative)
((Y2 case) = nominative)
```

Figure 7b: A rule that enforces agreement between an irrational plural noun as subject and an adjectival phrase as a predicate

7.2.3 Subject–verb number agreement

Recall that the Arabic verb is in singular if it precedes the subject. Therefore, in Arabic generation, it has to be decided whether to use the singular form of the Arabic verb and place it before the NP subject, or to use the number-agreeing form after the NP subject. This decision is taken when handling the sentence level, where it is already known whether the subject NP is pronominal or not, and the word order can be determined.

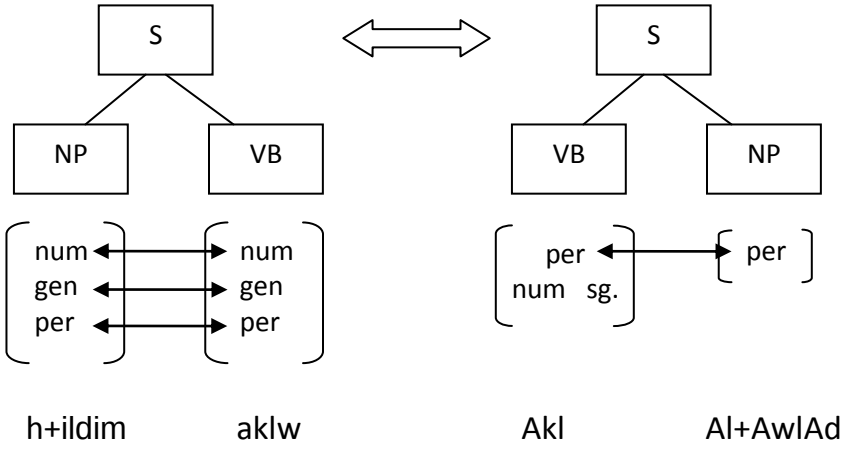

Figure 8a: A tree representation of a rule that enforces subject-verb agreement

```
{S_VB_NP_swap, 1}
;; SL: HLIDIM AKLW
;; TL: Akl AlAwlAd
S::S [NP VB] -> [VB NP]  # POS mapping
(X1::Y2)                 # POS alignment
(X2::Y1)
((X1 num) = (X2 num))    # Hebrew side agreement
((X1 gen) = (X2 gen))
((X1 per) = (X2 per))
((Y1 num) = singular)    # Arabic side agreement
((Y1 per) = (Y2 per))
```

Figure 8b: A rule that enforces subject-verb number agreement and word reordering when translating from Hebrew to Arabic

7.2.4 Dual number

When generating Hebrew, an explicit transfer of the Arabic dual number is needed. Since Arabic encodes dual number on verbs, nouns and adjectives, and Hebrew uses the plural form with the explicit cardinal number, this structure is created using designated rules. If the Arabic disambiguator determined that a verb or an adjective are dual, we generate the Hebrew plural form. When we encounter an Arabic dual noun, we

add an explicit cardinal number before the Hebrew noun in its plural form. The cardinal number has to agree on gender with the following noun (since Hebrew numerals are specified for gender), so two different rules add the cardinal number in the relevant gender: *šti* `two.fem' for feminine nouns, and *šni* `two.masc' for masculine nouns. When translating other morphemes that have to agree with the Hebrew noun on gender and number, such as verbs and adjectives, we convert them into the plural form, so the agreement in Hebrew holds with no further processing.

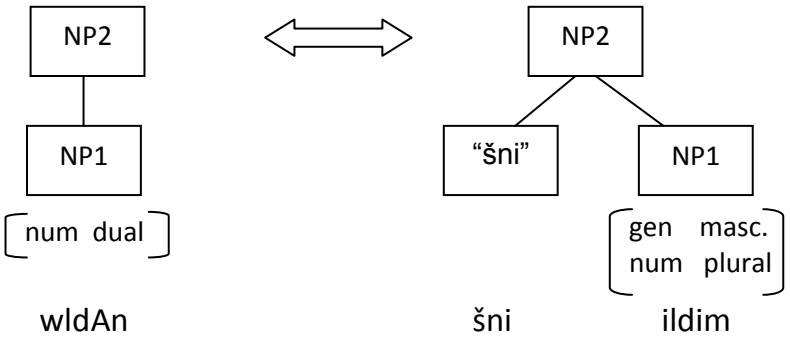

Figure 9a: A tree representation of a rule that maps Arabic dual number into the parallel Hebrew construction, using the explicit masculine cardinal number *šni* `two.masc'

```
{NP2_dual_masc,1}
;;SL: wldAn
;;TL: $NI ILDIM
NP2::NP2   [NP1]   ->   ["$NI"
NP1]
(X1::Y2)
((X1 num) = dual)
((Y2 gen) = masculine)
```

Figure 9b: A rule that maps Arabic dual number into the parallel Hebrew construction, using the explicit masculine cardinal number *šni* `two.masc'

7.2.5 Mood

Hebrew verbs in the future tense may be translated into the indicative imperfective and subjunctive imperfective forms in Arabic. As the choice of the proper mood is determined by the preceding Arabic word, transfer rules are perfectly placed to address the issue. If the preceding word is a preposition denoting intention, we choose the subjunctive form; otherwise, we choose the indicative form. This also reduces the lattice size. Negated Arabic verbs in the past tense also have two possible translations: the negated perfective form *mA ktbt* 'I didn't write', and the imperfective jussive form with the negative preposition *lm Aktb* 'I didn't write'. We generate both structures and let the LM choose according to local context. As for other usages of the imperfective jussive tense, these are rare cases that involve specific prepositions. Therefore these constructions are dealt with explicitly using designated transfer rules.

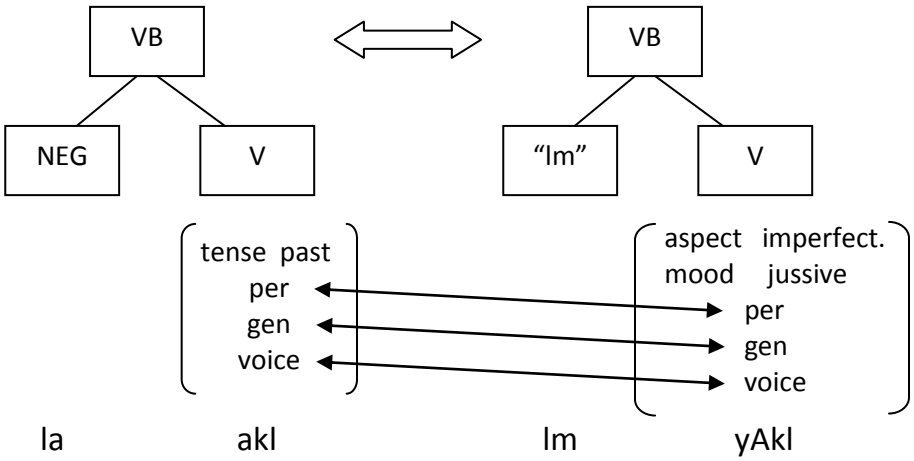

Figure 10a: A rule that maps Hebrew negated verb in past tense to Arabic negated verb in imperfective jussive mood, preceded by the negation particle *lm*

```
{VERB_NEG_lm,0}
;;SL: LA AKL
;;TL: lm yAkl
VB::VB [NEG V] -> ["lm" V]
(X2::Y2)
((X2 tense) = past)
((Y2 aspect) = imperfect)
((Y2 mood) = jussive)
((Y2 per) = (X2 per))
((Y2 gen) = (X2 gen))
((Y2 voice) = (X2 voice))
```

Figure 10b: A rule that maps Hebrew negated verbs in the past tense to Arabic negated verbs in imperfective jussive mood, preceded by the negation particle *lm*

8. Evaluation and error analysis

The two MT systems were fully implemented, although their coverage is still limited. To evaluate the performance of the systems, we created two test sets, one for each direction. All sentences in our development and test sets were extracted from newspaper texts; the Hebrew reference corpus was manually translated by two translators to Arabic, whereas for the Arabic reference corpus we obtained three human translations. In the Hebrew-to-Arabic test set, 84% of the Hebrew side morphemes had at least one entry in our bilingual lexicon, and 87% of the Arabic side morphemes in the Arabic-to-Hebrew test set had at least one such entry. As the systems are still not fully fledged, and several components are not yet functioning at full scale, we constrain the evaluation to smaller, simpler sentences for which we have good lexical coverage of the source language sentence. We selected all sentences of length 10 (words) or less, with at most one totally unknown morpheme in our lexicon. This resulted in a set of 39 sentences in the Hebrew-to-Arabic system, and 28 sentences in the Arabic-to-Hebrew system. Out-of-vocabulary (OOV) morphemes in the input sentences were manually supplemented for the smaller evaluation sets. As a baseline, we use the same systems with no grammar rules. Figure 11 below depicts actual translations produced by our systems compared to the baseline system on some of our development set sentences.

Consider (25): the translation using the grammar (25b) reflects correct transfer of number and enforcement of N-Adj agreement in both NPs (*mqwrršmi swdni* `official Sudani resource', and *šni Tiisim rwsim* `two Russian pilots'). In addition, the dual form in Arabic, which does not productively exist in Hebrew, is properly translated into the plural form in the noun, adjective and verb, and the explicit Hebrew numeral *šni* 'two' is generated in the correct gender. However, the passive form of the verbs is not properly generated, due to a mistake in determining the passive voice in the Arabic input. In the baseline system (25c), agreement is violated in both NPs, the dual number is not properly handled, and the input Arabic adjective *rwsyyn* ('Russian') is assigned the wrong POS in the output.

25.a AEln mSdr rsmy sdAny An TyAryn rwsyyn AxtTfA
inform.past.3ms source.sg official.sg sudani.sg that pilot.du russian.du kidnap.past.3.du
‘Official Sudani sources informed that two Russian pilots were kidnapped’ (Arabic input)

25.b mqwr ršmi swdni hwdi’ š šni Tiisim rwsim xTpw
source.sg official.sg sudani.sg inform.past.3ms that two pilot.m.pl russian.pl kidnap.pl
`Official Sudani sources informed that two Russian pilots kidnapped’ (Heb. with grammar)

25.c hwdi’w mqwrwt ršmiim swdni š Tiis rwsih xTwp
inform.past.3mp source.pl official.pl sudani.sg that pilot.m.sg Russia kidnap.sg.passive
`Informed official Sudani sources that a pilot Russia kidnapped’ (Hebrew without grammar)

26.a Akd AlHryry An ElAqt+h mE swryA ttTwr AijAbA
emphesize.past.3ms AlHariri that relation+his with Syria evolve.fut.3fs positively
`AlHariri confirmed that his relations with Syria are evolving positively' (Arabic input)

26.b hxriri ašr š+ qšr +w 'm swrih hštnh xiwbi
the+Hariri confirm.past.3ms that relation his with Syria change.past.3ms positive.m.sg
`The Hariri confirmed that his relations with Syria changed positive' (Hebrew with grammar)

26.c ašr hxriri š+ qšr awtw 'm swrih hštnh xiwbi
confirm.past.3ms the+Hariri that relation him with Syria change.past.3ms positive.m.sg
`Confirmed the Hariri that relation him with Syria changed positive' (Heb. without grammar)

27.a xbri+h šl h+nšiah hm mšpTnim m'wlim
friend.m.pl+her of the+president.f.sg they lawyer.m.pl excellent.m.pl
`The president's friends are excellent lawyers' (Hebrew input)

27.b AEDA' Al+r}ys hm mHAmwn mmtAzp
member.pl the+president.m.sg they lawyer.m.pl.nom excellent.m.sg.nom/gen
'The members of the president are excellent lawyer' (Arabic with grammar)

27.c Sdyq+y An Aly Al+r}ys+h mHAmyAF mmtAzAF
friend+my that to the+president+his lawyer.sg.acc excellent.sg.acc
incoherent (Arabic without grammar)

Figure 11: Translation examples: comparing our system with the grammar (b) to the system without the grammar (c), in Arabic-to-Hebrew direction (sentences 25,26) and Hebrew-to-Arabic (27)

In (26), the grammar-based system (26b) correctly generates the possessive pronoun (as opposed to (26c)). However, in both systems the proper name *AlHryry*, and the Arabic adverb *AyjAbA* `positively' were not properly translated. In (27), the grammar-based system (27b) correctly handles the Hebrew double genitive construction, translating it to the parallel Arabic *idafa* construction, and correctly treats the nominal predicate construction with the copula. There are still errors in N-Adj agreement on number and gender, and in the translation of the Hebrew subject noun *nšiah* (wrong gender). These issues arise from lattice explosion, since there were simply too many possibilities to explore. The baseline translation (27c), on the other hand, is totally incoherent.

We compared sample translations of our systems to those of Google Translate. Example (28b) demonstrates correct N-Adj agreement for rational and irrational plural nouns and correct treatment of NP conjunction structure. In example (28c), Google fails on generating the correct constituent structure, lexical translation of `policewomen' and enforcing agreement, and generates an incoherent result. Example (29b) demonstrates correct translation of the preposition, differing word order, V-Subj number agreement in Arabic, and conversion of a possessive construction using *šl* from Hebrew to *Idafa* construction in Arabic. In example (29c), Google fails on translating the Hebrew verb correctly, enforcing case, and the correct choice of preposition (*HDr* requires a direct object). Example (30b) demonstrates a correct translation of the double genitive and verbless predicate constructions. Google's translation in (30c) is incoherent.

28.a mkwniwt ipwt w+šwTrwt ipwt
car.pl.f pretty.pl.f and+policewomen.f pretty.pl.f
`Pretty cars and pretty policewomen' (Hebrew input)

28.b syArAt jmylp w+$rTyAt jmylAt
car.pl.f pretty.sg.f and+policewomen.f pretty.pl.f
`Pretty cars and pretty policewomen' (Stat-XFER)

28.c syArAt $rTp lTyf lTyf
car.pl.f police.sg.f pretty.sg.m pretty.sg.m
`The police's cars pretty pretty' (Google)

29.a h+ncigim šlkm nkxw b+išibh
the+representative.pl.m you.pl.m.poss attend.past.3.pl in+the+meeting.sg.f
`Your representatives attended the meeting' (Hebrew)

29.b HDr mmvlw+km Al+jlsp
attend.past.sg.m representative.pl.m.nom+your the+meeting
`Your representatives attended the meeting' (Stat-XFER)

29.c w+mmvlw+km Al+HADryn fy Al+AjtmAE
and+representative.pl.m.nom+your attend.ptcp.pl.m.def.acc/gen in the.meeting
`And your representatives that attended the meeting' (Google)

30.a mkwnit+h šl h+mnhlt gdwlh
car.sg.f+she.poss of the+principal.sg.f big.sg.f
`The principal's car is big'

30.b syArp Al+mdyrp kbyrp
car.sg.f the+principal.sg.f big.sg.f
`The principal's car is big' (Stat-XFER)

30.c Alf AlrAysy ll+syArAt
thousand main.sg.m.indef to+the+car.pl.f
`The cars' thousand main' (Google)

Following are automatic evaluation results on our simplified test set. Table 1 lists BLEU (Papineni et al., 2002) and METEOR (Lavie et al., 2004a) scores for both systems.

	With rules		Without Rules	
	BLEU	METEOR	BLEU	METEOR
Hebrew-to-Arabic	0.107	0.301	0.143	0.310
Arabic-to-Hebrew	0.275	0.46	0.231	0.417

Table 1: Evaluation results

Evidently, the Arabic-to-Hebrew system performs much better than the Hebrew-to-Arabic one. The grammar yields a significant improvement in the Arabic-to-Hebrew system, but it actually damages the Hebrew-to-Arabic system. The main reason for the deterioration in quality of translation using the grammar is lattice explosion, due to the great number of hypotheses. This is caused by two major factors:

1. Lacking a high-quality morphological disambiguator for Hebrew
2. The number of possibilities returned by the Arabic generator. When using a smaller bilingual lexicon with fewer translation options, the output is far better. We are currently working on ways to solve this issue, by incorporating a morphological disambiguator for Hebrew, and minimizing the number of results returned by the Arabic generator by merging results with identical surface forms and different feature structures.

To better understand these results, we performed a deep analysis of five sentences in each direction, focusing on the various potential sources of errors during the translation process. Table 2 lists the number of errors that can be attributed to each component: lexicon, grammar, decoder (when the correct hypothesis is present in the lattice but not selected), morphological analyzer, generator and disambiguation module.

	Lexicon	Grammar	Decoder	Analyzer	Generator	Disambiguation
H2A	14	11	4	3	1	
A2H	5	11	3		1	2

Table 2: Number of errors by type

We now take a closer look at one of the sentences in the smaller test set. The Arabic input is *nHn dA}mA nqwl lhm A*hbwA wqEwA AlAtfAq* ('we always tell them: go sign the agreement'), and the Hebrew references are

1. *anxnw tmid awmrim lhm lkw xtmw 'l hhskm*
2. *tmid awmrim lhm: lkw, xtmw 'l hhskm* (twice).

Our Arabic-to-Hebrew system produces the following output:

31. anxnw tmidi amr l+hm kli h+hskm ngn
    we constant tell.past.3ms to+them tool the+agreement play

Several errors occur in the translation of this sentence:

1. The Arabic lexical entry dA}mA is not matched. The reason is that our lexicon is specified for case diacritics, whereas the analyzer's output does not include them.
2. The pair *wqE* and *xtm* 'sign' is missing from our dictionary.
3. The Arabic disambiguation module wrongly chooses the verbal template *(A*ohab-*a instead of **ahab-a*), and predicts the wrong aspect (perfective instead of imperative).
4. The grammar lacks a rule for Subj-ADV-V. As a result, subject-verb agreement is not enforced.
5. The grammar lacks a rule for translating Arabic imperfective to Hebrew present tense.
6. The grammar lacks a rule that inserts the Hebrew preposition *'l* 'on'; there is no matching preposition in the Arabic input.

From the detailed error analysis and its numerical summary a clearer picture of the development status appears, where the grammar and lexicon are the crucial factors responsible for most of the errors. While augmenting and tuning the rules in the grammar is relatively easy, augmenting the bilingual lexicon is a hard task that currently remains open.

9. Translating prepositions

9.1 The challenge

As previously mentioned, translating prepositions is a complicated problem[8]. Prepositions have only a vague semantic meaning, yet they are often critical to the coherence of the output text. Prepositions are often strongly related (both semantically and distributionally) to the verb they attach to. In many cases, the usage of prepositions is idiomatic, and can vary in synonymous verbs even in the same language.

1. *hkh at ~ hrbic l-* `hit' (Hebrew)

2. *rah at ~ hstkl ‘l / b-* ~ cph b- `see, watch, look' (Hebrew)

3. *tfrj ElY ~ nZr AlY ~ lAHZ* (dir. obj.) `look, watch, notice' (Arabic)

4. *tfAdY (dir. obj.) / mn ~ A$fq En ~ tHrz mn* `watch out from, be careful of' (Arabic)

5. *think of ~ ponder* (dir. obj.) ~ think about (English)

In addition, the mapping between adjunct-preceding prepositions (such as temporals, locatives or adverbial phrases) may depend on the lexical content of the following phrase. For example, the Hebrew preposition *b* `in, at' is used before different types of locative, temporal and instrumental adjuncts, while the English preposition changes in these cases. In the following example it is translated into four different English prepositions (`on’, ‘at’, ‘in’ and null [empty preposition]).

32. dbrti **b+** iwm xmiši **b+** ‘rb **b+** Tlpwn **b+** šwq **b+** irwšlim

talk.1sg.past in day fifth in evening in phone in the.market in jerusalem

`I talked on Thursday evening on the phone at the market in Jerusalem' (Hebrew)

As a result, prepositions cannot be translated word-to-word. Common phrase-based SMT does not give a solution to this problem, since the coherence of the output text is

[8] Throughout this discussion the direct object is treated as a preposition (marked here by the symbol *at* in Hebrew).

modeled only be an n-gram language model. Even though such a language model can help in local contexts to discriminate in favor of the correct preposition, in long-distance dependencies the LM cannot help.

9.2 Possible solutions

Possible solutions to the aforementioned challenges may include:

1. Acquiring accurate and comprehensive statistics from a very large monolingual corpus of the target language regarding the distribution of verbs with their subcategorized prepositions and the NP-head following the prepositions. For example, such statistics could tell that the triples (*talked, to, him*) and (*talk, about, him*) are common, but (*talk, on, him*) would be very rare. As a backoff model, one could use a bigram model of the prepositions and the following NP-head, e.g., (on, Wednesday). This may help in the case of temporal and locative adjuncts that are less related to the preceding verb. Once these data are acquired, they may be used in the process of scoring hypotheses, if a parser is incorporated in the process.

One major shortcoming of this approach is the difficulty of acquiring these data. Another problem is the ability to generalize from these statistics: if "John" does not appear in the corpus, there would be no information regarding (*talk, to, John*) nor regarding (*talk, on, John*). The backoff to a bigram model would not help in these cases, and what is desired is the ability to represent the thematic roles of the arguments, and the selectional restrictions on these arguments (e.g., [+human]). In addition, there has to be an available high quality parser for the target language, and it should be incorporated during the decoding step, which is a heavy burden on performance.

2. Acquiring lexical and semantic mapping between verbs, the type of their arguments, the selectional restrictions on them, and the possible prepositions used for this relation. This can be done using a mapping from surface words to lexical ontologies, like WordNet (Miller, 1995), and to a syntactic-semantic mapping like VerbNet (Kipper et al., 2000) which lists the relevant preceding preposition. Similar work has been done by Shi and Mihalcea (2005) for the purpose of semantic parsing. These lexical-semantic

resources can help map between the verb and its possible arguments with their thematic roles, including selectional restrictions on them (expressed lexically, using a WordNet synset, like human or concrete).

The dominant shortcoming of this problem is that such explicit lexical and semantic resources exist mainly for English. Ongoing projects aim at expanding these resources to further languages, but they are currently not available. In addition, even when translating into English, this information can only assist in limiting the number of possible prepositions but not in determining them. For example, one can talk *on* the event, *after* the event, or *at* the event. The information that can disambiguate this question is in the source sentence.

3. Allowing translation of source-language prepositions to a limited set of possible target-language prepositions, and then using both target-language constraints on possible verb-preposition matches and an n-gram language model to choose the most adequate solution. Despite the fact that this solution does not model the probability of the target preposition given its verb and the original sentence, it limits the number of possible translations by taking into account the target-language verb and the possible constraints on the prepositions it licenses. This method is also the most adequate for implementing in a statistical decoder, such as the one used in Stat-XFER. We chose to implement this method in our system.

9.3 Translating prepositions between Hebrew and Arabic

Hebrew and Arabic share many similar prepositions such as *b*+ `in, at, with' and *l*+ `to'. However, there are prepositions that exist only in one of the languages, such as *En* `on, about' (Arabic). The different usage of prepositions between the two languages is significant and common. In example (33), the Arabic preposition *En* should be translated into the Hebrew direct object (the definite accusative marker *at*). However, example (34) is the opposite case where the Arabic direct object (no preposition) should be translated into the Hebrew preposition *b*+.

33.a AErb Al+wzyr En Aml+h
express.past.3ms the+minister on hope+his
`The minister expressed his hope' (Arabic)

33.b h+šr hbi' at tqwt +w
ha+sar hibi'a et tiqvato
the+minister express.past.3ms acc.def. hope+his
`The minister expressed his hope' (Hebrew)

34.a HDr Al+wzyr Al+jlsp
attend.past.3ms the+minister the+meeting
`The minister attended the meaning' (Arabic)

34.b h+šr nkx b+ h+išibh
ha+sar naxax b+ a+yeshiva
the+minister attend.past.3ms in the+meeting
`The minister attended the meaning' (Hebrew)

Here we see that despite the lexical and semantic similarity between many Hebrew and Arabic prepositions, their licensing by a semantically-equivalent verb is different in both languages.

An important issue is the selection of prepositions to model. Since there is no need to map each preposition in Arabic to each preposition in Hebrew, we focused on a small list of the common prepositions in both languages. We empirically counted prepositions in monolingual corpora from the news domain in both languages. The Arabic corpus size is 500K tokens, while the Hebrew corpus size is 120K tokens. Not surprisingly, the most common prepositions were those that are commonly used before complements. Results are listed in figure 12.

	prep	count	pct	acc_pct
1	fy	13128	18.7	18.7
2	dir. obj.	12626	17.9	36.7
3	l	9429	13.4	50.1
4	b	7253	10.3	60.4
5	mn	6859	9.7	70.2
6	ElY	5304	7.5	77.8
7	AlY	4458	6.3	84.1
8	En	1871	2.6	86.8
9	mE	1380	1.9	88.8
10	byn	1045	1.4	90.3

	prep	count	pct	acc_pct
1	b	6030	31.6	31.6
2	l	3386	17.7	49.3
3	dir.obj.	3250	17.0	66.3
4	m	1330	6.9	73.3
5	‘l	1066	5.5	78.9
6	k	354	1.8	80.7
7	‘m	338	1.7	82.5
8	am	200	1.0	83.6
9	bin	191	1.0	84.6
10	‘d	159	0.8	85.4

Figure 12: Counts of Arabic (left) and Hebrew (right) most common prepositions collected on monolingual corpora from the news domain. Arabic data is based on 70K prepositions, which comprise 14% of the corpus. Hebrew data is based on 19K prepositions, which comprise 16% of the corpus. The columns represent the lexical prepositions, their count in the corpus, the percentage out of all prepositions, and the accumulated percentage with all the higher-ranked prepositions[9].

Based on this evidence, we decided to focus on the set of first 9 Arabic prepositions (*fy, l-, b-, mn, ElY,AlY, En, mE* and the direct object), and the 6 Hebrew prepositions (*b-, l-, m-, ‘l, ‘m*, and the direct object).[10] These are the most common complement-preceding prepositions too, and therefore pose the main challenge for the task of MT.

9.4 Implementation

We implemented the last method for our Arabic-to-Hebrew system in two phases. In the first phase, a monolingual resource for Hebrew was created, which gave a statistical score for each Hebrew verb and its possible prepositions[11]. There was no distinction

[9] We ignore the Hebrew genitive marker *$l,* collapse the prepositions *m-* and *mn* together, and count direct objects whether they are marked with the indefinite accusative marker *at* or not.

[10] The decision not to include the Hebrew preposition k- stems from the fact that it has a direct Arabic counterpart k-, and that there is very little diversity in translation here.

[11] We would like to thank Hanna Fadida for sharing this resource.

between complement-preceding prepositions and adjunct-preceding ones. We used all the verb-preposition pairs whose score was higher than a certain threshold, set empirically, giving a total of 1402 verbs and 2325 verb-preposition pairs.

In the second stage, this information was incorporated into the Stat-XFER system in two different components: The Hebrew morphological generator as a feature for each such verb, and the grammar as constraints between the verb and its possible prepositions.

The Hebrew generator was modified to return an additional feature, allowed_preps, for each verb with such relevant information. For example, the Hebrew verb *sipr* `tell' had the feature of:

(allowed_preps = (*OR* at l))

Whenever the Hebrew generator returns a form of the verb *sipr*, the feature allowed_preps contains the possible prepositions *at* (dir. obj.) and *l* `to', that are licensed by this verb.

In addition, the grammar was modified to enforce constraints between the verb and its object. This was done by adding a new node, OBJ, which accounts for both direct objects and indirect objects. The strings of the prepositions are propagated as a feature to the OBJ node (see figure 13a below), and then in the sentence-level rule this feature is checked against the allowed_preps feature of the verb (figure 13b).

The first rule maps Arabic NP to Hebrew NP and marks the preposition *at* on the Hebrew OBJ node as the feature prep. The middle rule maps Arabic PP to Hebrew PP, and marks the Hebrew PP (referred to here as *Y1 lex*) on the Hebrew OBJ node as the feature prep. The rule at the bottom maps an Arabic NP to a Hebrew PP starting with the preposition *b* `in, at'.

```
{OBJ_ACC_AT,0}
;;SL: Al+ ktAb
;;TL: AT H+ SPR
OBJ::OBJ [NP] -> ["AT" NP]
(X1::Y2)
((X1 def) = +)
((Y2 prep) = AT)      # marking prepositions
(X0 = X1)
```

```
{OBJ_PP,0}
;;SL: mn Al+ $ms
;;TL: M H $M$
OBJ::OBJ [PREP NP] -> [PREP NP]
(X1::Y1)
(X2::Y2)
((Y0 prep) = (Y1 lex)) # marking prepositions
(X0 = X1)
(Y0 = Y1)
```

```
{OBJ_NP_PP_B, 0}
;;SL: (HDr) Al+ jlsp
;;TL: (NKX) B I$IBH
OBJ::OBJ [NP] -> ["B" NP]
(X1::Y2)
((Y0 prep) = B)       # marking prepositions
(X0 = X1)
(Y0 = Y2)
```

Figure 13a: propagating the surface form of the preposition as a feature of the OBJ node

```
{S_VB_NP_OBJ_swap, 1}
;; SL: HDr alr}ys Aljlsp
;; TL: HN$IA NKX BI$IBH
S::S [VB NP OBJ] -> [NP VB OBJ]
(X1::Y2)
(X2::Y1)
(X3::Y3)
((X1 num) = singular)              # Arabic side agreement
((X1 per) = (X2 per))
((Y1 num) = (Y2 num))              # Hebrew side agreement
((Y1 gen) = (Y2 gen))
((Y1 per) = (Y2 per))
((Y2 allowed_preps) = (Y3 prep)) # Hebrew preposition
```

Figure 13b: Enforcing agreement between VB and OBJ on the Hebrew side in the sentence level rule

Following are some examples from the system's output, which illustrate the behavior of the system. There are four types of syntactic mappings between Arabic and Hebrew arguments: (NP, NP), (NP, PP), (PP, NP) and (PP, PP). Sentences (35) and (36) demonstrate correct translation of the Arabic direct object into the Hebrew direct object (with and without the definite accusative marker *at*, respectively). Sentence (37) demonstrates the correct translation of the Arabic direct object to a Hebrew PP with the preposition *l-*. Sentence (38) demonstrates the correct translation of an Arabic PP to a Hebrew direct object, and sentence (39) demonstrates the translation of Arabic PP starting with *b-* `in, with' into a Hebrew PP with *'m* `with'.

35.a rAyt Al+wld
saw.past.1s the+boy
`I saw the boy' (Arabic, NP object)

35.b raiti at h+ild
saw.past.1s acc.def the+boy
`I saw the boy' (Hebrew, definite NP object)

36.a rAyt wldA
saw.past.1s boy.acc.indef
`I saw a boy' (Arabic, NP object)

36.b raiti ild
saw.past.1s boy
`I saw a boy' (Hebrew, indefinite NP object)

37.a Drb Al+Ab Al+wld
hit.past.3ms the+father the+boy
`The father hit the boy' (Arabic, NP)

37.b h+ab hrbic l+ h+ild
the+father hit.past.3ms to the+boy
`The father hit the boy' (Hebrew, PP object)

38.a AErb Al+wzyr En Aml+h
express.past.3ms the+minister on hope+his
`The minister expressed his hope' (Arabic, PP object)

38.b h+šr hbi' at tqwt+w
the+minister express.past.3ms acc.def. hope+his
`The minister expressed his hope' (Hebrew, definite NP object)

39.a AjtmE Al+wzyr b+ Al+wld
meet.past.3ms the+minister in the+boy
'The minister met the boy' (Arabic, PP object)

39.b h+šr npgš 'm h+ild
the+minister meet.past.3ms with the+boy
'The minister met the boy' (Hebrew, PP object)

In (35), the input Arabic NP is definite and identified as accusative case. In a designated rule, we add the string *at* before the corresponding Hebrew output, to mark the definite direct object. We create a node of type OBJ for both (direct) objects, with the feature *prep* representing the lexical content of the preposition in the target language. Finally, in the sentence level rule, we validate that the Hebrew verb licenses a direct object, by unifying the *prep* feature of OBJ with the *allowed_preps* feature of VB. In (36), a similar process occurs, but this time no additional token of *at* is added. The same preposition *at* is marked as a feature of OBJ (since it marks the direct object), and again, the *prep* feature of OBJ is validated against the *allowed_preps* feature of VB. Example (37) is created using a rule that maps Arabic direct object to a Hebrew PP starting with a different preposition, here *l-* `to'. There is such a rule for every Hebrew preposition, since we have no prior knowledge of which prepositions should be generated. We mark the lexical preposition *l-* on the prep feature of the Hebrew OBJ node, and again, this is validated in the sentence level against the prepositions allowed by VB. In example (38) we use rules that map Arabic PP to Hebrew NP, in which the Arabic preposition is not translated at all, and the Hebrew definite accusative marker *at* is added according to the definiteness of the Hebrew NP. The only difference in example (39) compared to previous examples is the translation of the Arabic preposition into a different Hebrew preposition. This was done in the bilingual lexicon, in a lexical entry that mapped Arabic *b-* `in, with*'* to Hebrew *'m* `with'.

All of these rules help in expanding the lexical variety of the prepositions on one hand (like in example (39)), while disqualifying some hypotheses with prepositions that are not licensed by the preceding verb (which results in ungrammatical sentences) by using unification-style constraints. After this process, there may still be different hypotheses in the lattice, from which the decoder statistically chooses the best one.

9.5 Evaluation

For evaluation of our treatment for prepositions, we used the same set of 28 short sentences (up to 10 surface words) that we to used to evaluate our Arabic-to-Hebrew system (all the input and output sentences are listed in Appendix I). As a baseline system, we used exactly the same setup, except for two differences:

1. We omitted the restrictions on which prepositions Hebrew verbs license, such that each verb can be followed by each preposition.
2. We limited the lexical variance between prepositions in the lexicon, to only allow the common translations of prepositions. For example, we omitted the mapping of *ElY* `on' (Arabic) → *b* `with' (Hebrew), but we left the mapping of *ElY* `on' (Arabic) → *'l* `on' (Hebrew).

Table 3 details the automatic scores for each system.

	BLEU	METEOR
with prepositions	0.370	0.560
w/o prepositions	0.325	0.526

Table 3: Automatic evaluation scores for translating prepositions.

As can be seen, the system with the special treatment for prepositions outperforms the baseline system. While analyzing the results, we saw that the baseline system incorporated prepositions that are not licensed by the preceding verb, and the LM did not help in choosing the hypothesis with the correct preposition, if such a hypothesis was generated at all. Example (40) is a good example of the difference between the outputs of both systems.

40.a Akd AlHryry ElY AltzAm +h b+ Al+byAn
emphasize.past.3ms AlHaryry on obligation+his in the+announcement

Al+wzAry l+ Hkwmp Al+whdp Al+wTnyp
the+ministerial to government the+unity the+national

`Alharyry emphasized his obligation in the ministerial announcement to the national government' (Arabic input)

40.b alxriri hdgiš at xwbt +w b+ h+ hwd'h
Alharyry emphasize.past.3ms def.acc obligation+his in+the+announcement

h+mmšltit l+ mmšlt h+axdwt h+lawmit
the+governmental to+government the+unity the+national

`Alharyry emphasized his obligation in the governmental announcement to the national government' (Hebrew output, with prepositions handling)

40.c alxriri aišr 'l zkiwn šl+w b +h+ hwd'h
Alharyry confirm.past.3ms on permit of+his in+the+announcement

h+mmšltit l+ mmšlt h+axdwt h+lawmit
the+governmental to+government the+unity the+national

`Alharyry confirmed on his permit in the governmental announcement to the national government' (incoherent Hebrew output, with no prepositions handling)

In (40b), the verb *hdgiš* `emphsized' is followed by the correct definite accusative marker *at*. In (40c), the verb *aišr* `approved' is followed by a wrong preposition `*l* `on', which is not licensed in this location. After that, the lexical selections for the translations were different and not as fluent as in (40b), and the output is only partially coherent.

9.6 Future directions

One possible direction to continue the research on this topic is incorporating more information about different types of arguments a verb may receive, such as complemental clause (in our case, with the Hebrew relativizer *š* `that'), or an infinite verb. Another direction could be acquiring and incorporating such information on verb-derived nouns, which license the same prepositions as the parallel verbs. For example, *xtimh 'l hskm* `signing.noun an agreement', where the Hebrew preposition *'l* `on' must be used, as in a similar verbal from, like *xtm 'l hskm* `singed an agreement'. However, this may lead to problematic issues like PP attachment, since we do not know if the PP (here *'l hskm* 'on an agreement') relates to a noun or to a preceding verb.

10. Summary

In this work we surveyed the problem of Machine Translation between morphologically-rich and resource-poor languages, such as Hebrew and Arabic. We listed main linguistic similarities and differences between the two related languages, and discussed the implications of the challenges. We offered possible solutions to these challenges, and implemented the approach of using direct translation utilizing linguistic resources. We implemented two MT systems between Hebrew and Arabic for both directions. The implementation included developing a grammar that maps linguistic constructions between source and target languages, and incorporating into our systems existing resources such as morphological analyzers, disambiguators and generators. We gave automatic evaluation results, and proved that for Arabic-to-Hebrew using the grammar yields better output than the baseline system which did not use the grammar. We conducted an error analysis of the output, and concluded that the grammar and lexicon are responsible for the lion's share of errors in our translations.

As a linguistically-interesting and complex question, we focused on the problem of translating prepositions, which involves great variety between languages and often includes non-local dependencies. We focused on a closed set of 9 Arabic prepositions and 6 Hebrew prepositions (including the direct object, which is sometimes not overt). We implemented our solution in the Arabic-to-Hebrew systems. We used monolingual data extracted from Hebrew corpora which mapped verbs to possible prepositions. We incorporated this information into our grammar, which helped us constrain and control the translation of prepositions. Finally, we gave successful empirical results from this experiment.

Future plans

There are many possible directions to continue the work presented above.

1. Continue to improve the grammar and the lexicon in both directions. Since the lexical and syntactic coverage are always sub-optimal (specifically in our case), this may well lead to better translations.
2. Move to *ktiv male* in the Hebrew side of the lexicon. As mentioned in Sections 4.1.1 and 5.1 on Hebrew spelling, Hebrew has no single uniform convention of writing, and the Hebrew side of our bilingual lexicon does not reflect spelling forms of common text words. Successfully adding or changing from *ktiv xaser* into *ktiv male* Hebrew writing in our lexicon may be the difference in many cases between successful translation and an out-of-vocabulary word.
3. Continue research on transfer of prepositions according to preceding verb. For example, enhancing the monolingual knowledge we have on target language argument structure for both verbs and nouns, while incorporating this as constraints into the system.
4. Improve the ability of the system to use multi-word expressions during the process of hypotheses generation and decoding.
5. Incorporate transliteration from Arabic to Hebrew (or in the opposite direction) for unknown words.

Appendix I

Attached are the 28 sentence-pairs that were used to evaluate the Arabic-to-Hebrew system for the task of translating prepositions. The sentences appear in original language writing. The Hebrew output in some cases is perfect (including a correct full parse of both input and output sentences), and in the vast majority of the cases comprehensible. Agreement and voice are sometimes wrong, when the parse or morphological analysis of the input sentence are wrong, and the output sentence is a concatenation of several hypotheses.

1. ابدى وزير الخارجية استعداده للقاء قيادة حركة المقاومة الاسلامية "حماس"
   שר הביע החוץ נכונותו פגישה של הנהגת תנועת ההתנגדות האיסלמית "חמאס"
2. نحن دائما نقول لهم اذهبوا وقعوا الاتفاق
   אנחנו תמידי אומר להם אסיר חתם ההסכם
3. أعلن مصدر رسمي سوداني أن طيارين روسيين اختطفا
   מקור רשמי סודני הודיע ששני טייסים רוסים חטפו
4. يحذر عباس من فشل المفاوضات
   עבאס מזהיר כשלון במשא ומתן
5. الاعتراف بإسرائيل دولة للشعب اليهودي هو ضروري
   ההכרה בישראל מדינה לעם היהודי הוא הכרחי
6. شدد الرئيس الفلسطيني على أن المفاوضات ستبحث قضايا الوضع النهائي
   הנשיא הפלסטיני הדגיש בשהמשא ומתן יעסקו בשאלה של המעמד הסופי
7. الأراضي الفلسطينية المحتلة عام 1967 هي السبب الرئيسي للنزاع
   הארץ הפלסטיניים הכבושים שנת 1967 הוא הגורם הראשי למאבק
8. لا حل سحريا للصراع
   אין פתרון מקסים למאבק
9. يذهبون الفلسطينيون إلى هذه المفاوضات بعد أن فرض الأمر عليهم
   הולך הפלסטיני למשא ומתן הזה אחרי שהדבר הטיל עליהם
10. لن نسمح بمرور قافلة شريان الحياة
    לא נרשה תנועת שיירה של עורק החיים
11. جدير بالذكر أن القافلة انطلقت أول أمس السبت
    מתאימה את הזוכר שהשיירה פרצה בראשון אתמול של השבת

12. اشتبك أعضاء قافلته مع الشرطة المصرية ووقعت إصابات بين الطرفين
חבר שיירתו התעמת עם המשטרה המצרית ופגיעות נפלו בשני הצדדים
13. أكد الحريري على التزامه بالبيان الوزاري لحكومة الوحدة الوطنية
חרירי הדגיש את חובתו בהודעה הממשלתית לממשלת האחדות הלאומית
14. أكد الحريري أن علاقته مع سوريا تتطور إيجابا ً
חרירי הדגיש שקשורו עם סוריה מתפתחת חיוב
15. لن يسمح بتمرير الفتنة
בשום אופן לא מרשה העברה של המבחן
16. قتل رئيس الوزراء السابق رفيق الحريري
ראש השר הקודם רפיק חרירי הרג
17. وجهت إسرائيل رسالة إلى روسيا
ישראל הפנתה איגרת לרוסיה
18. نحن نعلم بالصفقة منذ فترة وأجريت محادثات مع الروس
אנחנו יודעים את ההסכם מאז תקופה וערך שיחה עם הרוסי
19. للأسف تتم الصفقة على مراحل
לצער ההסכם יתבצע בשלב
20. فإن الصفقة مقلقة جدا لإسرائيل
אכן שההסכם מדאיג מאוד לישראל
21. تحاول إسرائيل أن تعرقل صفقة السلاح
ישראל משתדלת שהסכם של הנשק יפריע
22. أعلنت إسرائيل صراحة أنها ستحاول إفشال هذه الصفقة
ישראל הודיעה בהירות שהוא הכשלה תשתדל ההסכם הזה
23. يجب ضمان استمرار التفوق الإسرائيلي النوعي في المنطقة لسنين طويلة
צריך ערבות של התמדה של היתרון הישראלי האיכותי באזור לשנים ארוכות
24. قتل 5 أشخاص بينهم 3 مسلمين
הורג 5 איש בהם 3 מוסלמים
25. مع مرور الوقت بدأ المنتخب العراقي يمسك زمام المبادرة
עם חלוף הזמן התחיל הנבחרת העיראקית תיקח את רסן היזמה
26. المنتخب الفلسطيني بحث عن هدف يعيد الأمور إلى نصابها
הקבוצה הפלסטינית חיפשה שער הדבר ישיב לכנה
27. استعاد المنتخب الفلسطيني حضوره على أرض الملعب
הקבוצה הפלסטינית החזיר את נוכחותו בארץ האיצטדיון
28. إحرز كريم هدفين وإستهم في الهدف الثالث
כרים כבשו שני שערים והשתתפו אותם במטרה השלישי

References

Bogdan Babych, Anthony Hartley, and Serge Sharoff. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of MT Summit XI*, Copenhagen, Denmark. 2007.

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

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311, 1993

Tim Buckwalter. 2002. Buckwalter transliteration. http://www.qamus.org/transliteration.htm

Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, Philadelphia, 2004.

John Chandioux. METEO: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public. 1976. META 21:127-133.

David Chiang. 2006. An introduction to synchronous grammars. Tutorial available at http://www.isi.edu/~chiang/papers/synchtut.pdf

Noam Chomsky. 1956. Three models for the description of language. Information Theory, IEEE Transactions 2 (3).

Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan A. Pérez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, and Kepa Sarasola. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of the 10th European Association for Machine Translation Conference*, pages 79-86, 2005.

Ahmet Cüneyd Tantuğ, Eşref Adali and Kemal Oflazer, Machine translation between turkic languages. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 189-192, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Nissim Francez and Shuly Wintner. Unification Grammars. Cambridge: Cambridge University Press (Forthcoming)

Ahmed El Kholy and Nizar Habash. Techniques for Arabic morphological detokenization and orthographic denormalization. In *Proc. of LREC-2010.*

Nizar Habash. 2004. Large scale lexeme based Arabic morphological generation. In *Proc. of Traitement Automatique du Langage Naturel (TALN-04).*

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proc. of HLTNAACL*.

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An improved statistical transfer system for French–English machine translation. In *Proc. Of StatMT '09: the Workshop on Statistical Machine Translation*.

Jan Hajic, Petr Homola, and Vladislav Kubo. A simple multilingual machine translation system. In *Proceedings of the MT Summit IX*, 2003.

Alon Itai and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98.

Karin Kipper, Hoa Trang Dang, and Martha Palmer Class-based construction of a verb lexicon. In *Proceedings of Seventeenth National Conference on Artificial Intelligence AAAI*, 2000 (Austin,TX, July 2000).

Philipp Koehn and Hieu Hoang, Factored translation models. In *Proc. EMNLP+CoNLL*, pages 868-876, Prague, 2007.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proc. of EMNLP-CoNLL*.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitj´os, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. 2003. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. ACM Transactions on Asian Language Information Processing (TALIP), 2(2):143–163.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. The significance of recall in automatic metrics for mt evaluation. In R. E. Frederking and K. Taylor, editors, AMTA, volume 3265 of Lecture Notes in Computer Science, pages 134–143. Springer, 2004a. ISBN 3-540-23300-8.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 2004b.

Alon Lavie. 2008. Stat-XFER: A general search-based syntax driven framework for machine translation. In Alexander F.Gelbukh, editor, CICLing, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375. Springer.

Wolfgang Lörscher. *Translation Performance, Translation Process, and Translation Strategies*: *A Psycholinguistic Investigation*. Tübingen, Gunter Narr, 1991.

George A. Miller. 1995. WORDNET: A Lexical Database for English. Communications of ACM(11): 39-41.

Christian Monson, Ariadna Font Llitjós, Vamshi Ambati, Lori Levin, Alon Lavie, Alison Alvarez, Roberto Aranovich, Jaime Carbonell, Robert Frederking, Erik Peterson, and Katharina Probst. Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), may 2008. ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/

Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120105775299168.

Kazunori Muraki. 1987. PIVOT: Two-phase machine translation system. In *MT Summit Manuscripts and Program*, pages 81–83.

Uzzi Ornan. 2003. The Final Word. University of Haifa. Press, Haifa, Israel. In Hebrew.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In *Proceedings of the HLT-NAACL Workshop on Syntax and Structure in Statistical Translation*, Rochester, NY, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135

Florence Reeder, Measuring MT Adequacy Using Latent Semantic Analysis. In Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts, pages 176-184. 2006.

Lei Shi and Rada Mihalcea. Putting the pieces together: Combining FrameNet, VerbNet, and WordNet for robust semantic parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico. 2005.

Peter Toma. SYSTRAN as a multilingual machine translation system. In *Commission of the European Communities. Overcoming the language barrier* (München, Vlg. Dokumentation, 1977), 569-581

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proc. of RANLP*.

Bernard Vauquois, A Survey of Formal Grammars and Algorithms for Recognition and Translation in Machine Translation. FIP Congress-68, Edinburgh, pp. 254-260, 1968.

Shuly Wintner. 2005. Unification: Computational issues. In *Keith Brown, editor, Encyclopedia of Language and Linguistics, volume 13*, pages 238–250. Elsevier, Oxford, second edition.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. Of ACL*.

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In Proceedings *of ACL 39*, 2001.