A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora

Arul Menezes and Stephen D. Richardson

Microsoft Research
One Microsoft Way
Redmond, WA 98008, USA
arulm@microsoft.com
steveri@microsoft.com

Abstract

Translation systems that automatically extract transfer mappings (rules or examples) from bilingual corpora have been hampered by the difficulty of achieving accurate alignment and acquiring high quality mappings. We describe an algorithm that uses a best-first strategy and a small alignment grammar to significantly improve the quality of the transfer mappings extracted. For each mapping, frequencies are computed and sufficient context is retained to distinguish competing mappings during translation. Variants of the algorithm are run against a corpus containing 200K sentence pairs and evaluated based on the quality of resulting translations.

1 Introduction

A machine translation system requires a substantial amount of translation knowledge typically embodied in bilingual dictionaries, transfer rules, example bases, or a statistical model. Over the last decade, research has focused on the automatic acquisition of this knowledge from bilingual corpora. Statistical systems build translation models from this data without linguistic analysis (Brown, 1993). Another class of systems, including our own, parses sentences in parallel sentence-aligned corpora to extract transfer rules or examples (Kaji, 1992) (Meyers, 2000) (Watanabe, 2000). These systems typically obtain a predicate-argument or dependency structure for source and target sentences, which are then aligned, and from the resulting alignment, lexical and structural translation correspondences are extracted, which are then represented as a set of rules or an example-base for translation.

However, before this method of knowledge acquisition can be fully automated, a number of issues remain to be addressed. The alignment and transfer-mapping acquisition procedure must acquire rules with very high precision. It must be robust against errors introduced by parsing and sentence-level alignment, errors intrinsic to the corpus, as well as errors resulting from the alignment procedure itself. The procedure must also produce transfer mappings that provide sufficient context to enable the translation system utilizing these mappings to choose the appropriate translation for a given context.

In this paper, we describe the alignment and transfer-acquisition algorithm used in our machine translation system, which attempts to address the issues raised above. This system acquires transfer mappings by aligning pairs of *logical form structures (LFs)* similar to those described by Jensen (1993). These LFs are obtained by parsing sentence pairs from a sentence-aligned bilingual corpus. (The problem of aligning parallel corpora at the sentence level has been addressed by Meyers (1998b) Chen (1993) and others and is beyond the scope of this paper).

We show that alignment using a best-first strategy in conjunction with a small alignment grammar improves the alignment and the quality of the acquired transfer mappings.

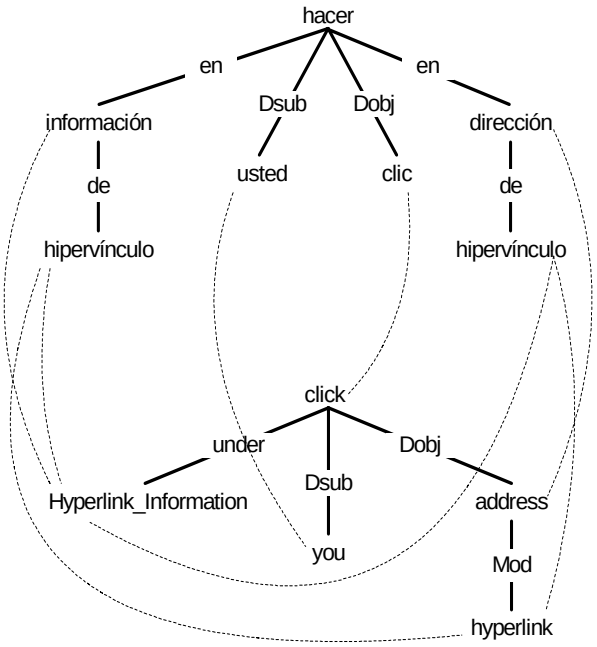

Figure 1a: Lexical correspondences

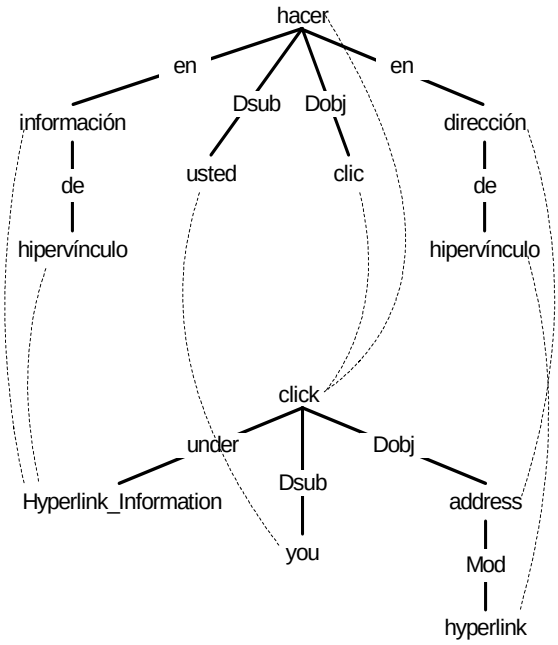

Figure 1b: Alignment Mappings

2 Logical Form

A Logical Form (LF) is an unordered graph representing the relations among the most meaningful elements of a sentence. Nodes are identified by the lemma of a content word and directed, labeled arcs indicate the underlying semantic relations. Logical Forms are intended to be as independent as possible of specific languages and their grammars. In particular, Logical Forms from different languages use the same relation types and provide similar analyses for similar constructions. The logical form abstracts away from such language-particular aspects of a sentence as constituent order, inflectional morphology, and certain function words.

Figure 1a illustrates the LFs for the following Spanish sentence and its corresponding English translation, which we use in example below.

En Información del hipervínculo, haga clic en la dirección del hipervínculo.

Under Hyperlink Information, click the hyperlink address.

3 Alignment

We consider an alignment of two logical forms to be a set of mappings, such that each mapping is between a node or set of nodes (and the relations between them) in the source LF and a node or set of nodes (and the relations between them) in the target LF, where no node participates in more than one such mapping. In other words, we allow one-to-one, one-to-many, many-to-one and many-to-many mappings but the mappings do not overlap.

Our alignment algorithm proceeds in two phases. The first phase establishes tentative lexical correspondences between nodes in the source and target LFs. The second phase aligns nodes based on these lexical correspondences as well as structural considerations. The algorithm starts from the nodes with the tightest lexical correspondence (“best-first”) and works outward from these anchor points.

We first present the algorithm, and then illustrate how it applies to the sentence-pair in Figure-1.

3.1 Finding tentative lexical correspondences

We use a bilingual lexicon that merges data from several sources (CUP, 1995), (SoftArt, 1995), (Langenscheidt, 1997), and inverts target-to-source dictionaries to improve coverage. Our Spanish-English lexicon contains 88,500 translation pairs. We augment this with 19,762 translation correspondences acquired using statistical techniques described by Moore (2001).

Like Watanabe (2000) and Meyers (2000), we use a lexicon to establish initial tentative word correspondences. However, we have found that even a relatively large bilingual dictionary has only moderately good coverage for our purposes. Hence, we pursue an aggressive matching strategy for establishing tentative word correspondences. Using the bilingual dictionary together with the derivational morphology component in our system (Pentheroudakis, 1993), we find direct translations, translations of morphological bases and derivations, and base and derived forms of translations. Fuzzy string matching is also used to identify possible correspondences. We have found that aggressive over-generation of correspondences at this phase is balanced by the more conservative second phase and results in improved overall alignment quality.

We also look for matches between components of multi-word expressions and individual words. This allows us to align such expressions that may have been analyzed as a single lexicalized entity in one language but as separate words in the other.

3.2 Aligning nodes

Our alignment procedure uses the tentative lexical correspondences established above, as well as structural cues, to create affirmative node alignments. A set of alignment grammar rules licenses only linguistically meaningful alignments. The rules are ordered to create the most unambiguous alignments ("best") first and use these to disambiguate subsequent alignments. The algorithm and the alignment grammar rules are intended to be applicable across multiple languages. The rules were developed while working primarily with a Spanish-English corpus, but have also been applied to other language pairs such as French, German, and Japanese to/from English.

The algorithm is as follows:

1. Initialize the set of unaligned source and target nodes to the set of all source and target nodes respectively.
2. Attempt to apply the alignment rules in the specified order, to each unaligned node or set of nodes in source and target. If a rule fails to apply to any unaligned node or set of nodes, move to the next rule.
3. If all rules fail to apply to all nodes, exit. No more alignment is possible. (Note: some nodes may remain unaligned).
4. When a rule applies, mark the nodes or sets of nodes to which it applied as aligned to each other and remove them from the lists of unaligned source and target nodes respectively. Go to step 2 and apply rules again, starting from the first rule.

The alignment grammar currently consists of 18 rules. Below we provide the specification for some of the most important rules.

1. *Bidirectionally unique translation*: A set of contiguous source nodes S and a set of contiguous target nodes T such that every node in S has a lexical correspondence with every node in T and with no other target node, and every node in T has a lexical correspondence with every node in S and with no other source node. Align S and T to each other.
2. *Translation + Children*: A source node S and a target node T that have a lexical correspondence, such that each child of S and T is already aligned to a child of the other. Align S and T to each other.
3. *Translation + Parent*: A source node S and a target node T that have a lexical correspondence, such that a parent P_s of S has already been aligned to a parent P_t of T. Align S and T to each other.
4. *Verb+Object to Verb:* A verb V_1 (from either source or target), that has child O that is not a verb, but is already aligned to a verb V_2, and either V_2 has no unaligned parents, or V_1 and V_2 have children aligned to each other. Align V_1+O to V_2.
5. *Parent + relationship:* A source node S and a target node T, with the same part-of-

speech, and no unaligned siblings, where a parent P_s of S is already aligned to a parent P_t of T, and the relationship between P_s and S is the same as that between P_t and T. Align S and T to each other.

6. *Child + relationship:* Analogous to previous rule but based on previously aligned children instead of parents.

Note that rules 4-6 do not exploit lexical correspondence, relying solely on relationships between nodes being examined and previously aligned nodes.

3.3 Alignment Example

In this section, we illustrate the application of the alignment procedure to the example in Figure 1. In the first phase, using the bilingual lexicon, we identify the lexical correspondences depicted in Figure-1a as dotted lines. Note that each of the two instances of *hipervínculo* has two ambiguous correspondences, and that while the correspondence from *Información* to *Hyperlink Information* is unique, the reverse is not. Note also that neither the monolingual nor bilingual lexicons have been customized for this domain. For example, there is no entry in either lexicon for *Hyperlink_Information*. This unit has been assembled by general-purpose "Captoid" grammar rules. Similarly, lexical correspondences established for this unit are based on translations found for its individual components, there being no lexicon entry for the captoid as a whole.

In the next phase, the alignment rules apply to create alignment mappings depicted in Figure-1b as dotted lines.

Rule-1: *Bidirectionally unique translation*, applies in three places, creating alignment mappings between *dirección* and *address*, *usted* and *you*, and *clic* and *click*. These are the initial "best" alignments that provide the anchors from which we will work outwards to align the rest of the structure.

Rule-3: *Translation + Parent*, applies next to align the instance of *hipervínculo* that is the child of *dirección* to *hyperlink*, which is the child of *address*. We leverage a previously created alignment (*dirección* to *address*) and the structure of the logical form to resolve the ambiguity present at the lexical level.

Rule-1 now applies (where previously it did not) to create a many-to-one mapping between *información* and *hipervínculo* to *Hyperlink_Information*. The uniqueness condition in this rule is now met because the ambiguous alternative was cleared away by the prior application of Rule-3.

Rule-4: *Verb+Object to Verb* applies to rollup *hacer* with its object *clic*, since the latter is already aligned to a verb. This produces the many-to-one alignment of *hacer* and *clic* to *click*

4 Acquiring Transfer Mappings

Figure-2 shows the transfer mappings derived from the alignment example in Figure-1.

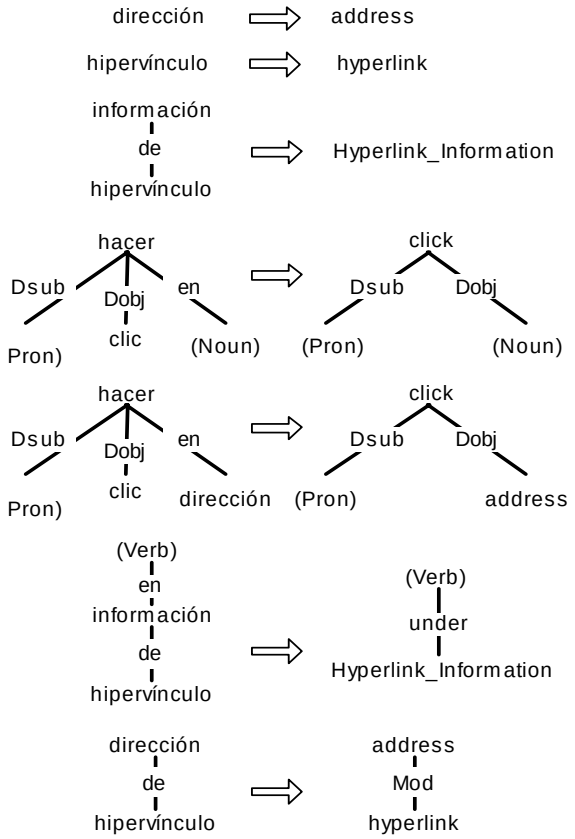

Figure-2 : Transfer mappings acquired

4.1 Transfer mappings with context

Each mapping created during alignment forms the core of a family of mappings emitted by the transfer mapping acquisition procedure. The alignment mapping by itself represents a

minimal transfer mapping with no context. In addition, we emit multiple variants, each one expanding the core mapping with varying types and amounts of local context.

We use linguistic constructs such as noun and verb phrases to provide the boundaries for the context we include. For example, the transfer mapping for an adjective is expanded to include the noun it modifies; the mapping for a modal verb is expanded to include the main verb; the mapping for a main verb is expanded to include its object; mappings for collocations of nouns are emitted individually and as a whole. Mappings may include "wild card" or under-specified nodes, with a part of speech, but no lemma, as shown in Figure 2.

4.2 Alignment Post-processing

After we have acquired transfer mappings from our entire training corpus, we compute frequencies for all mappings. We use these to resolve conflicting mappings, i.e. mappings where the source sides of the mapping are identical, but the target sides differ. Currently we resolve the conflict by simply picking the most frequent mapping. Note that this does not imply that we are committed to a single translation for every word across the corpus, since we emitted each mapping with different types and amounts of context (see section 4.1). Ideally at least one of these contexts serves to disambiguate the translation. The conflicts being resolved here are those mappings where the necessary context is not present.

A drawback of this approach is that we are relying on a priori linguistic heuristics to ensure that we have the right context. Our future work plans to address this by iteratively searching for the context that serves to optimally disambiguate (across the entire training corpus) between conflicting mappings.

4.2.1 Frequency Threshold

During post-processing we also apply a frequency threshold, keeping only mappings seen at least N times (where N is currently 2). This frequency threshold greatly improves the speed of the runtime system, with negligible impact on translation quality (see section 5.6).

5 Experiments and Results

5.1 Evaluation methodology

In the evaluation process, we found that various evaluation metrics of alignment in isolation bore very little relationship to the quality of the translations produced by a system that used the results of such alignment. Since it is the overall translation quality that we care about, we use the output quality (as judged by humans) of the MT system incorporating the transfer mappings produced by an alignment algorithm (keeping all other aspects of the system constant) as the metric for that algorithm.

5.2 Translation system

Our translation system (Richardson, 2001) begins by parsing an input sentence and obtaining a logical form. We then search the transfer mappings acquired during alignment, for mappings that match portions of the input LF. We prefer larger (more specific) mappings to smaller (more general) mappings. Among mappings of equal size, we prefer higher-frequency mappings. We allow overlapping mappings that do not conflict. The lemmas in any portion of the LF not covered by a transfer mapping are translated using the same bilingual dictionary employed during alignment, or by a handful of hard-coded transfer rules (see Section 5.7 for a discussion of the contribution made by each of these components). Target LF fragments from matched transfer mappings and default dictionary translations are stitched together to form an output LF. From this, a rule-based generation component produces an output sentence.

The system provides output for every input sentence. Sentences for which spanning parses are not found are translated anyway, albeit with lower quality.

5.3 Training corpus

We use a sentence-aligned Spanish-English training corpus consisting of 208,730 sentence pairs mostly from technical manuals. The data was already aligned at the sentence-level since it was taken from sentence-level translation memories created by human translators using a commercial translation-memory product. This data was parsed and aligned at the sub-sentence

level by our system, using the techniques described in this paper. Our parser produces a parse in every case, but in each language roughly 15% of the parses produced are "fitted" or non-spanning. Since we have a relatively large training corpus, we apply a conservative heuristic and only use in alignment those sentence-pairs that produced spanning parses in both languages. In this corpus 161,606 pairs (or 77.4% of the corpus) were used. This is a substantially larger training corpus than those used in previous work on learning transfer mappings from parsed data. Table-1 presents some data on the mappings extracted from this corpus using Best-First.

Total Sentence pairs	208,730
Sentence pairs used	161,606
Number of transfer mappings	1,202,828
Transfer mappings per pair	7.48
Num. unique transfer mappings	437,479
Num. unique after elim. conflicts	369,067
Num. unique with frequency > 1	58,314
Time taken to align entire corpus (on a 800MHz PC)	74 minutes
Alignment speed	35.6 sent/s

Table-1: Best-first alignment of training corpus

5.4 Experiments

In each experiment we used 5 human evaluators in a blind evaluation, to compare the translations produced by the test system with those produced by a comparison system. Evaluators were presented, for each sentence, with a reference human translation and with the two machine translations in random order, but not the original source language sentence. They were asked to pick the better overall translation, taking into account both content and fluency. They were allowed to choose "Neither" if they considered both translations equally good or equally bad.

All the experiments were run with our Spanish-English system. The test sentences were randomly chosen from unseen data from the same domain. Experiment-1 used 200 sentences and each sentence was evaluated by all raters. Sentences were rated better for one system or the other if a majority of the raters agreed. Experiments 2-4 used 500 sentences each, but each sentence was rated by a single rater.

In each experiment, the test system was the system described in section 5.2, loaded with transfer mappings acquired using the techniques described in this paper (hereafter "Best-First").

5.5 Comparison systems

In the first experiment the comparison system is a highly rated commercial system, Babelfish (http://world.altavista.com).

Each of the next three experiments varies some key aspect of Best-First in order to explore the properties of the algorithm.

5.5.1 Bottom Up

Experiment-2 compares Best-First to the previous algorithm we employed, which used a bottom-up approach, similar in spirit to that used by Meyers (1998a).

This algorithm follows the procedure described in section 3.1 to establish tentative lexical correspondences. However, it does not use an alignment grammar, and relies on a bottom-up rather than a best-first strategy. It starts by aligning the leaf nodes and proceeds upwards, aligning nodes whose child nodes have already aligned. Nodes that do not align are skipped over, and later rolled-up with ancestor nodes that have successfully aligned.

5.5.2 No Context

Experiment-3 uses a comparison algorithm that differs from Best First in that it retains no context (see section 4.1) when emitting transfer mappings.

5.5.3 No Threshold

The comparison algorithm used in Experiment-4 differs from Best First in that the frequency threshold (see section 4.2.1) is not applied, i.e. all transfer mappings are retained.

Comparison System	Num. sentences Best-First rated better	Num. sentences comparison system rated better	Num. sentences neither rated better	Net percentage improvement
Babelfish	93 (46.5%)	73 (36.5%)	34 (17%)	10.0%
Bottom-Up	224 (44.8%)	111 (22.2%)	165 (33%)	22.6%
No-Context	187 (37.4%)	69 (13.8%)	244 (48.8%)	23.6%
No-Threshold	112 (22.4%)	122 (24.4%)	266 (53.2%)	-2.0%

Table-2: Translation Quality

5.6 Discussion

The results of the four experiments are presented in Table-2.

Experiment-1 establishes that the algorithm presented in this paper automatically acquires translation knowledge of sufficient quantity and quality as to enable translations that exceed the quality of a highly rated traditional MT system. Note however that Babelfish/Systran was not customized to this domain.

Experiment-2 shows that Best-First produces transfer mappings resulting in significantly better translations than Bottom-Up. Using Best-First produced better translations for a net of 22.6% of the sentences.

Experiment-3 shows that retaining sufficient context in transfer mappings is crucial to translation quality, producing better translations for a net of 23.6% of the sentences.

Experiment-4 shows that the frequency threshold hurts translation quality slightly (a net loss of 2%), but as Table-3 shows it results in a much smaller (approx. 6 times) and faster (approx 45 times) runtime system.

	Num mappings	Translation speed (500 sentences)
Best-First	58,314	173s (0.34s/sent)
No-Threshold	359,528	8059s (17s/sent)

Table-3: Translation Speed (500 sentences)

5.7 Transfer mapping coverage

Using end-to-end translation quality as a metric for alignment leaves open the question of how much of the translation quality derives from alignment versus other sources of translation knowledge in our system, such as the bilingual dictionary, or the 2 hand-coded transfer rules in our system. To address this issue we measured the contribution of each using a 3264-sentence test set. Table-4 presents the results. The first column indicates the total number of words in each category. The next four columns indicate the percentage translated using each knowledge source, and the percentage not translated respectively.

As the table shows, the vast majority of content words get translated using transfer-mappings obtained via alignment.

Our alignment algorithm does not explicitly attempt to learn transfer mappings for pronouns, but pronouns are sometimes included in transfer mappings when they form part of the context that is included with each mapping (see section 4.1). The 31.89% of pronoun translations that the table indicates as coming from alignment fall into this category.

Our algorithm does try to learn transfer mappings for prepositions and conjunctions, which are represented in the Logical Form as labels on arcs (see Figure-1). Mappings for prepositions and conjunctions always include the nodes on both ends of this arc. These mappings may translate a preposition in the source language to a preposition in the target language, or to an entirely different relationship, such as direct object, indirect object, modifier etc.

As the table shows, the system is currently less successful at learning transfer mappings for prepositions and conjunctions than it is for content words.

As a temporary measure we have 2 hand-coded transfer rules that apply to prepositions, which account for 8.4% of such transfers. We intend for these to eventually be replaced by mappings learned from the data.

	Number of instances	Alignment	Dictionary	Rules	Not translated
Content words	21,245	93.50%	4.10%	0%	2.4%
Pronouns	2,158	31.89%	68.20%	0%	0%
Prepositions/Conjunctions	6,640	32.00%	59.70%	8.4%	0%

Table-4: Coverage of transfer mappings, dictionary & rules

6 Conclusions and Future Work

We proposed an algorithm for automatically acquiring high-quality transfer mappings from sentence-aligned bilingual corpora using an alignment grammar and a best-first strategy.

We reported the results of applying the algorithm to a substantially larger training corpus than that used in previously reported work on learning transfer mappings from parsed data.

We showed that this approach produces transfer mappings that result in translation quality comparable to a commercial MT system for this domain.

We also showed that a best-first, alignment-grammar based approach produced better results than a bottom-up approach, and that retaining context in the acquired transfer mappings is essential to translation quality.

We currently rely on a priori linguistic heuristics to try to provide the right context for each transfer mapping. In future work, we plan to use machine-learning techniques to determine the extent of the context that optimally disambiguates between conflicting mappings.

References

Peter Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, 1993. "The mathematics of statistical machine translation" *Computational Linguistics, 19:263-312*

Cambridge University Press (1995), McCarthy, M. ed., *Cambridge Word Selector*

Stanley F. Chen, 1993. "Aligning sentences in bilingual corpora using lexical information" *Proceedings of ACL 1993*

Karen Jensen, 1993. "PEGASUS: Deriving argument structures after syntax." In *Natural Language Processing: The PLNLP Approach*. Kluwer Academic Publishers, Boston, MA.

Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto, 1992. "Learning Translation Templates from Bilingual Text" *Proceedings of COLING 1992*

Langenscheidt Publishers 1997, *The Langenscheidt Pocket Spanish Dictionary*

Adam Meyers, Roman Yangarber, Ralph Grishman, Catherine Macleod, and Antonio Moreno-Sandoval, 1998a. "Deriving transfer rules from dominance-preserving alignments", *Proceedings of COLING 1998*

Adam Meyers, Michiko Kosaka and Ralph Grishman, 1998b. "A multilingual procedure for dictionary-based sentence alignment" *Proceedings of AMTA 98*

Adam Meyers, Michiko Kosaka and Ralph Grishman, 2000. "Chart-based transfer rule application in machine translation" *Proceedings of COLING 2000*

Robert C. Moore 2001, "Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words" *Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001*

Joseph Pentheroudakis and Lucretia Vanderwende 1993, "Automatically identifying morphological relations in machine-readable dictionaries" *Ninth Annual conference of the University of Waterloo Center for the new OED and Text Research*

Stephen D. Richardson, William Dolan, Monica Corston-Oliver, and Arul Menezes 2001, "Overcoming the customization bottleneck using example-based MT", *Workshop on Data-Driven Machine Translation, ACL 2001*

SoftArt Inc (1995) *Soft-Art translation dictionary. Version 7*

Hideo Watanabe, Sado Kurohashi, and Eiji Aramaki, 2000. "Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation" *Proceedings of COLING 2000*