Translating with Examples

Andy Way,
School of Computer Applications,
Dublin City University,
Dublin, Ireland.

Email: away@compapp.dcu.ie

Abstract

Machine Translation (MT) systems based on Data-Oriented Parsing (DOP: Bod, 1998) and LFG-DOP (Bod & Kaplan, 1998) may be viewed as instances of Example-Based MT (EBMT). In both approaches, new translations are processed with respect to previously seen translations residing in the system's database. We describe the DOT models of translation (Poutsma 1998; 2000) based on DOP. We demonstrate that DOT1 is not guaranteed to produce the correct translation, despite provably deriving the most probable translation. The DOT2 translation model solves most of the problems of DOT1, but suffers from limited compositionality when confronted with certain data. Notwithstanding the success of DOT2, any system based purely on trees will ultimately be found wanting as a general solution to the wide diversity of translation problems, as certain linguistic phenomena require a description at levels deeper than surface syntax. We then show how LFG-DOP can be extended to serve as a novel hybrid model for MT, LFG-DOT (Way, 2001), which promises to improve upon DOT and other EBMT systems.

1 Problems for EBMT

One of the major advantages which is often claimed of EBMT is that the overall quality of translation increases incrementally as the set of stored translations increases. For example, Mima *et al.* (1998) report that in their EBMT system, translation quality rose in an almost linear fashion, from 30% with 100 examples to 65% with all 774 examples. They also note that there seems to be a limit beyond which adding further examples does not improve the overall translation quality.

While the chances of finding an exact match become greater as the corpus size increases, there are two knock-on effects whose impact on the EBMT system should be minimized. Firstly, adding more examples has a computational cost, especially if the examples need to be parsed: some EBMT systems (e.g. Sato & Nagao, 1990; Watanabe, 1994) store examples as annotated trees, for instance. Whether this is the case or not, adding more examples causes more storage problems, and adds to the complexity of the search and retrieval stages of the EBMT process. It is, therefore, unclear that one of the purported advantages of EBMT, namely a lessening in computational cost, is indeed a real benefit. Secondly, adding more examples may not be useful in practice. For example, a newly added translation pair may be identical to, or overlap other examples. Where a system involves the computation of a 'similarity metric' (e.g. Somers *et al.*, 1994), this may be influenced by the frequency of examples, so that the score attached to certain matches increases if a large number of similar examples are found. Alternatively, in other systems, identical or similar examples may just be redundant. Somers (1999:121) observes that "in such systems, the examples can be seen as surrogate 'rules', so that, just as in a traditional rule-based MT system, having multiple examples (rules) covering the same phenomenon leads

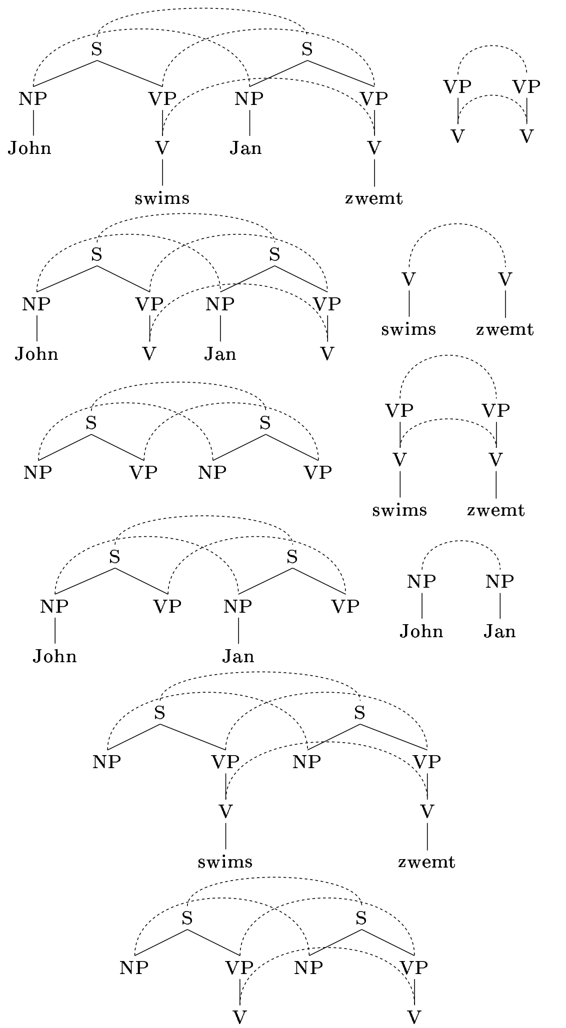


Figure 1: The complete DOT treebank for the linked sentence pair ⟨*John swims*, *Jan zwemt*⟩

to over-generation".

The two main problems for EBMT are *boundary definition* and *boundary friction*. The first of these describes the scenario where retrieved fragments may not be well-formed constituents. This is a particular problem for pure EBMT systems, where syntactic well-formedness needs to be ensured without grammatical information actually being employed. The second, boundary friction, is a problem in the retrieval process, in that context may not be taken into account. This may be illustrated by attempting to translate *I have a big dog* into German. An EBMT system might retrieve the close matches in (1):

(1) a. A big dog eats a lot of meat ⟶Ein großer Hund frisst viel Fleisch.

b. I have two ears ⟶Ich habe zwei Ohren.

From these examples, the useful translation fragments in (2) might be isolated by our EBMT system:

(2) a. A big dog ⟶Ein großer Hund

b. I have ⟶Ich habe

In this case, these fragments would be combined to give the translation *Ich habe ein großer Hund*, which is ungrammatical as we have an NP bearing nominative case in object position.

We shall show how DOT and LFG-DOT fare with these two main problems for EBMT. With strict notions of decomposition of fragments, the problem of boundary definition is unproblematic. DOT, however, may suffer from the problem of boundary friction, while given the additional syntactic information available in the Lexical-Functional Grammar (LFG: Kaplan & Bresnan, 1982) f-structures, this problem is considerably reduced in LFG-DOT.

2 Data-Oriented Translation

Poutsma (1998; 2000) has developed a model of translation based on Tree-DOP—Data-Oriented Translation (DOT). There are two different versions of DOT. We shall describe these briefly.

2.1 DOT1

Poutsma's DOT1 model (1998; 2000) is based on the methodology of Tree-DOP (cf. Bod, 1998), and relates POS-fragments between two languages (English and Dutch here), with an accompanying probability. DOT1 is parameterized on similar lines to Tree-DOP. Its **representations** are *linked* phrase-structure trees. Figure 1 shows the complete treebank for the linked sentence pair ⟨*John swims*, *Jan zwemt*⟩.[1]

[1] Here, and in future examples, we 'translate' proper names purely in order to differentiate completely source and target representations and strings.

These trees are augmented to incorporate semantic information, as a DOT1 treebank links semantically equivalent trees. It can be seen that the top left tree pair in Figure 1 represents the full parse trees for this translation pair. This tree pair is subjected to the DOT1 **decomposition** process (based on the DOP *Root* and *Frontier* fragmentation operations—essentially *Root* allows new tree fragments to be built by selecting a node to be the root node of a new tree, and deleting all other nodes except this new root and all nodes dominated by it, while *Frontier* selects a (possibly empty) set of nodes in the newly created subtree, excluding the root, and deletes all subtrees dominated by these selected nodes) so that the other tree pairs in Figure 1 are derived. Note also here that all S-rooted linked structures are derived via *Frontier*, while all others are produced via *Root*. It is these operations which delimit the boundary definitions in the DOT translation models.

Tree pair fragments are subjected to the **recombination** process in DOT using the same composition operation, namely leftmost substitution (to ensure the uniqueness of each derivation), as Tree-DOP. The target derivations are then assembled by replacing all subtrees of the source derivation by their linked target subtrees. Any pair of fragments which can legitimately combine with others results in a well formed derivation given the corpus. This causes a particular problem of boundary friction in DOT (cf. Figure 3) which is avoided in LFG-DOT (cf. (12)). Finally, the **probability** of target trees that are candidate translations is then calculated using the Tree-DOP probability model based on their relative frequencies in a treebank such as Figure 1. Importantly, therefore, DOT treebanks are *bags* of fragments, rather than sets. Accordingly, where duplication of 'similar' examples may be an impediment in other example-based systems, in DOT (and LFG-DOT) they are essential, as an increase in frequency of particular fragments directly contributes to the weight (in terms of probability) of the derivations in which they are involved. The probability of the parse tree of a particular translation is calculated by summing the probability of all possible derivations resulting in that parse tree. If different parse trees result for a particular translation, the respective probabilities of each with respect to the others can be calculated with respect to the corpus. Similarly, if different translations are presented as candidate target strings, their respective probabilities are given. As will will see with respect to (13)–(18) below, this is useful as certain MT systems cannot prevent the output of multiple translation candidates, even where some of these may be ungrammatical target strings. In such cases, an expert user is required to sift through these output strings to select the 'best' translation. This situation is avoidable where translations are output with probabilities, as in DOT and LFG-DOT, as translations may be ranked automatically (or pruned, if only the highest ranked translation is required).

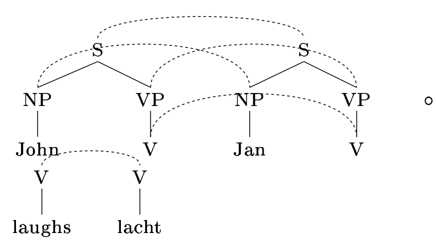
Given this trivial treebank, only the one translation pair can be processed. If we assume that the treebank fragments for the translation pair *Peter laughs* $\Longleftrightarrow$ *Piet lacht* are added to Figure 1 (these new fragments would be identical except for the different lexical material on the leaves of the tree fragments), then the two new translations in (3) can be handled on the basis of fragments already in the database:

(3) a. *John laughs* $\longleftrightarrow$ *Jan lacht*

b. *Peter swims* $\longleftrightarrow$ *Piet zwemt*

For example, one possible derivation of (3a) is that in (4):

(4)

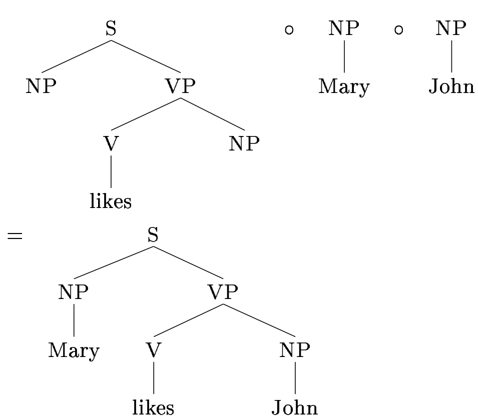


That is, the V nodes in the lower tree pair can be substituted at the appropriate $\langle$*source, target*$\rangle$ V nodes in the upper tree pair. The full parse trees for the two sentences ensue, resulting in a *bona fide* translation given this treebank. The two translations in (3) will have a slightly smaller probability than the two original translation pairs given the presence of their full linked parse trees in the treebank, each of which has a probability of $\frac{1}{12}$. That is, there is one occurrence of the actual parse trees themselves, and there are 12 trees with root `cat=s`. However, given the trivial nature of the examples, there will be no other translation candidates output so each translation will have a probability of 1. This will, of course, not always be the case with more complex examples. For instance, in section 4 we examine in two experiments the situation where multiple translations are output in LFG-DOT with accompanying probabilities given the existence of both 'default' and 'specific' fragments in the database.

2.1.1 A Problem for DOT1

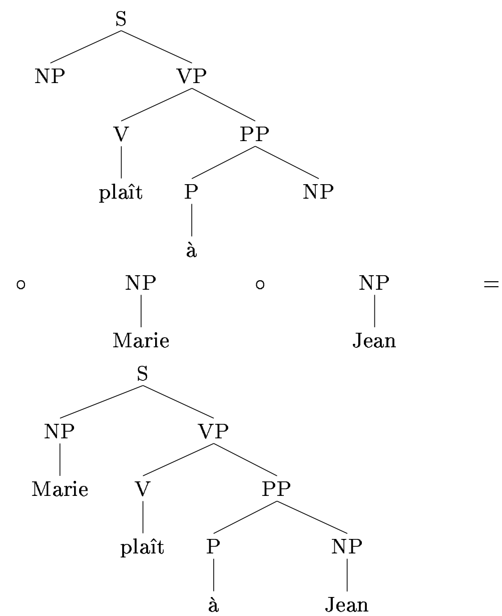
The DOT1 composition operation, based on leftmost substitution in the source trees, does not deal properly with translation cases where the word order differs significantly between two languages. Such an example is the *like* $\longleftrightarrow$ *plaire* relation-changing case. As soon as derivation (5) of the source sentence is arrived at, the desired linking of the English SUBJ with the French prepositional OBJ, and that of the English OBJ with the French SUBJ, are overridden by the composition operation of DOT1:

(5)



In this case the wrong translation of (5) is derived, as in (6):

(6)



It would appear that the adherence to leftmost substitution in the target given *a priori* leftmost substitution in the source is too strictly linked to the linear order of words, so that as soon as this deviates to any significant extent even between similar languages, DOT1 has a huge bias in favour of the incorrect translation. Even if the correct, non-compositional translation is achievable in such circumstances, it is likely to be so outranked by other wrong alternatives that it will be dismissed, unless all possible translations are maintained for later scrutiny by the user.

2.2 DOT2: An Improved Model of Translation

Poutsma overcomes this problem in DOT2 by redefining the composition operation of DOT1 to operate on *pairs* of trees, rather than on single trees. This new definition of composition ensures, among other things, that relation-changing cases such as *like* $\longleftrightarrow$ *plaire* are handled correctly: instead of the wrong derivation (6) of the source (5), we now obtain the correct translation *Jean plaît à Marie.* In DOT1, leftmost substitution in the source overrides the desired DOT-links between the French subject NP and the English object, as well as those between the French prepositional object and the English subject, given that composition is defined on the source tree *only*. Once pairs of trees are taken into account in DOT2, these links ensure the correct translation. Way (2001) shows that while DOT2 is able to handle certain 'hard' cases correctly, other examples, such as headswitching, are dealt with in a 'semi-compositional' manner. Consider the data in (7):

(7) a. DE: Johannes schwimmt gerne ⟷ EN: John likes to swim.

b. DE: Josef läuft zufällig ⟷ EN: Joseph happens to run.

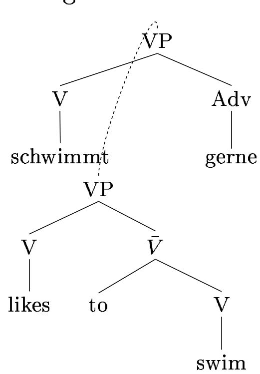
Presupposing the derivation of a monolingual treebank constructed from the German examples in (7), with two different NPs, verbs and adverbs, eight sentences are possible and can receive analyses with associated probabilities with respect to that DOP corpus. However, only four of these possible sentences can receive translations in a DOT corpus, namely the examples in (7) as well as those in (8), by simple substitution of the alternate NPs into the respective subject slots:

(8) a. DE: Josef schwimmt gerne ⟷ EN: Joseph likes to swim.

b. DE: Johannes läuft zufällig ⟷ EN: John happens to run.

The other four possible sentence pairs cannot be handled at all in a DOT treebank built from analyses of the strings in (7). This is due to the fact that the linked VP pairs are not broken down any further than at the root level. The contrast can be seen by examining the *schwimmt gerne* VP and its constituent DOP-fragments in Figure 2 and (9), which contains the single DOT linked VP pair:

(9)



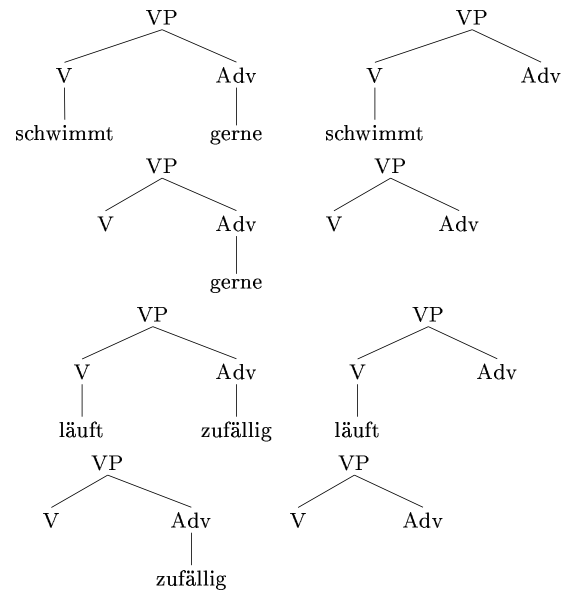


Figure 2: The monolingual DOP VP-fragments for a treebank built from the German examples in (7)

We cannot draw a link between *schwimmt* and *swim* in (9) as they are not translationally equivalent: one is a finite verb and the other an infinitive. We cannot, therefore, express the basic translation relations as ⟨*gerne, likes to*⟩ and ⟨*zufällig, happens to*⟩. Given this restriction, the only way that the other sentence pairs can be handled is if there is some prior linked pair *läuft gerne* ⟷ *likes to run* as well as a prior instance of the linked pair *schwimmt zufällig* ⟷ *happens to swim*. This is because these linked VP pairs are handled non-compositionally in DOT2 between German and English, but the monolingual VPs are treated compositionally in DOP. Contrast this situation with a DOT treebank designed to translate these 8 German strings into Dutch. Our starting point could be the German strings in (7) with their Dutch translations, as in (10):

(10) a. DE: Johannes schwimmt gerne ⟷ NL: Johan zwemt graag.

b. DE: Josef läuft zufällig ⟷ NL: Josef loopt toevallig.

Given these simple transfer (i.e. 'word for word') examples, a DOT treebank would resemble much more closely the monolingual DOP treebanks from which it is derived for the respective sentences in (10), as *every* constituent part of the German strings corresponds exactly to a constituent part of the Dutch strings. In the DOT treebank these links are made explicit. When we have a headswitching case, however, it is apparent that both DOT models would translate the sentences correctly *iff* prior examples of linked headswitching VPs exist in the treebank. Such translations would receive extremely low probabilities with respect to the corpus in the normal case as they are built with a minimal degree of compositionality (substitution of subject NP, no other derivations being possible). As these examples only ever occur rarely, the chances of DOT2 managing to translate these in practice becomes significantly lower than might otherwise be expected, as we require not only the presence of the adverb, but also its occurrence to be correlated exactly with the verb in question for translation to succeed. We shall show that the LFG-DOT4 model of translation is able to provide generalized translation fragments which enable fully compositional translation in these cases, as required.

2.3 Boundary Friction in DOT

It is clear that DOT2 is an improvement on DOT1. DOT1 cannot always explicitly relate parts of the source language structure to the corresponding, correct parts in the target structure, so fails to translate correctly where source and target strings differ significantly with respect to word order. In DOT2, correct translations are obtained along with some possible wrong alternatives. Notwithstanding the improved composition operation and probability model of DOT2, its ability to achieve the correct translation is compromised by a lesser amount of compositionality in the translation process. Given the small number of fragments playing a role in the derivation of some translations involving complex phenomena, almost the exact linked sentence pair may need to be present in order for a translation to be possible.

A further problem is that Poutsma's DOT models cannot distinguish ill-formed from well-formed

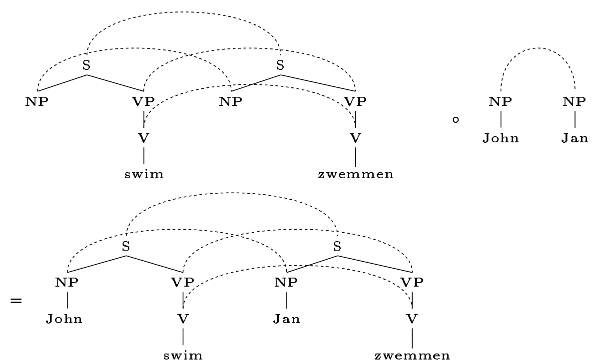


Figure 3: The Boundary Friction Problem in DOT

input. For example, both DOT models of translation would permit the derivations in Figure 3. That is, with no stipulation on subject-verb agreement, it is perfectly legitimate in DOP-based models (and EBMT systems based purely on trees, such as Watanabe 1992, as well, of course, as systems where examples are stored as pairs of strings, e.g. Somers *et al.*, 1990) to combine a singular subject with a plural verb and end up with a well-formed object. In DOT, we end up with a translation which is well-formed given the corpus. This example illustrates that while the presence of category information alleviates the problem of boundary friction to a certain extent (compared to 'pure' example-based methods, for instance, where the examples are stored as strings with no linguistic information present at all), this continues to be problematic for DOT in certain circumstances. We show in (12) that the analogous derivation in LFG-DOT would be deemed ungrammatical. That is, as soon as grammatical information is available via the accompanying f-structures, such a combination is impossible given the clash in NUM values for the subject and verb. Such ill-formed input can still be handled by relaxing grammaticality constraints such as these via *Discard*, but such translation pairs will be regarded as ungrammatical with respect to the corpus given their derivation via *Discard*; in DOT models, we have no such distinction.

Finally, of course, translation systems which are based purely on PS-trees will ultimately not be able to handle certain linguistic phenomena. DOP-based approaches are necessarily limited to those contextual dependencies actually occurring in the corpus, which is a reflection of surface phenomena only. Purely context-free models are insufficiently powerful to deal with all aspects of human language. Lexical Functional Grammar is known to be beyond context-free. It can capture and provide representations of linguistic phenomena other than those occurring at surface structure. With this in mind, the functional structures of LFG have been allied to the techniques of DOP to create a new model, LFG-DOP (Bod & Kaplan, 1998), which adds a measure of robustness (both with respect to unseen as well as ill-formed input) not available to models based solely on LFG.

3 LFG-DOT Models of Translation

Way (2001) presents four possible LFG-DOT translation models, all of which use LFG-DOP as their language models. LFG-DOP models are defined using the same four parameters as in Tree-DOP. Its **representations** are lifted *en bloc* from LFG theory, so that each string is annotated with a c-structure, an f-structure, and a mapping ϕ between them. Since we are now dealing with $\langle c, f\rangle$ pairs of structure, the *Root* and *Frontier* **decomposition** operations of DOP need to be adapted to stipulate exactly which c-structure nodes are linked to which f-structure fragments, thereby maintaining the fundamentals of c- and f-structure correspondence. Given that LFG c-structures are little more than annotated PS-trees allows us to proceed very much on the same lines as in Tree-DOP.

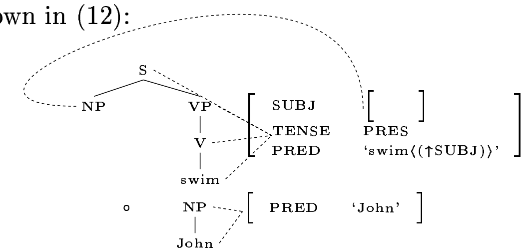
Root erases all nodes outside of the selected node, and in addition deletes all ϕ-links (informally, parts of the f-structure linked to a c-structure node, cf. the dotted lines in (12), for example) leaving the erased nodes, as well as all f-structure units that are not ϕ-accessible from the remaining nodes (for instance, features such as TENSE are deleted when the main verb is removed, as this feature—in English, at least—is inextricably linked to the verbal PRED). *Frontier* operates as in Tree-DOP, deleting all subtrees of the selected frontier nodes. Furthermore, it deletes all ϕ-links of these removed nodes together with any semantic form corresponding to the same nodes. Finally, a third, new operation, *Discard*, provides generalized fragments from those derived via *Root* and *Frontier* by freely deleting any combination of attribute-value pairs from an f-structure except those that are ϕ-linked to some remaining c-structure node, or that are governed by the local predicate. Its introduction also necessitates a new definition of the grammaticality of a sentence *with respect to a corpus*, namely any sentence having at least one derivation whose fragments are produced only by *Root* and *Frontier* and not by *Discard*. **Composition** is also a two-step operation. C-structures are combined by leftmost substitution, as in Tree-DOP, subject to the matching of their nodes. F-structures corresponding to these nodes are then recursively unified, and the resulting f-structures are subjected to the grammaticality checks of LFG. Finally, the **probability model** of LFG-DOP is again based on the relative frequency of a fragment, which in this case is a $\langle c, f\rangle$ pair. Bod & Kaplan give definitions of three possible competition sets from which fragments are sampled, depending on which point during the sampling stage the well-formedness conditions (the *Root* matching condition of DOP, or the Uniqueness, Completeness and Coherence conditions of LFG) are invoked. Note that given the non-monotonic property of the Completeness check, this can only be enforced after all other validity sampling has taken place.

3.1 LFG-DOT Model 1: Translation via τ

The first two LFG-DOT models propose the use of LFG's τ-equations to relate translation fragments between languages, the second in combination with γ, the function that links DOT source and target subtree fragments. Using τ-equations overcomes some of the problems of the DOT1 translation model. For instance, the LFG-MT solution (11) to the *like* $\longleftrightarrow$ *plaire* relation-changing case can be availed of quite straightforwardly:

$$\text{(11)} \quad \begin{array}{ll} \textit{like}: & (\tau\!\uparrow \text{PRED}) = \text{plaire} \\ & \tau(\uparrow \text{SUBJ}) = (\tau\!\uparrow \text{OBL}) \\ & \tau(\uparrow \text{OBJ}) = (\tau\!\uparrow \text{SUBJ}) \end{array}$$

That is, the subject of *like* is translated as the oblique argument of *plaire*, while the object of *like* is translated as the subject of *plaire*. In addition, DOP adds robustness to LFG-MT. LFG-DOT models of translation contain two monolingual LFG-DOP language models, so *Discard* can be run on both source and target sides. This means that LFG-DOT can cope with ill-formed or previously unseen input which LFG-MT would not be able to handle at all. Suppose that *John swim* is encountered as the source string, as in Figure 3. In LFG-DOP, this can only be interpreted if *Discard* relaxes certain constraints in the f-structures. One such derivation is shown in (12):

(12) 

That is, the `SUBJ:NUM:PL` path will have been relaxed in the sentential f-structure, with the `NUM = SG` feature removed from the lower NP f-structure. The NP c-structure can be substituted at the NP node in the upper c-structure, and the f-structures unified. The resultant f-structure would be input into the translation phase, which in LFG-DOT1 is quite simply the LFG-MT τ function. Taking the f-structure resulting from the derivation in (12) as input, the appropriate τ-equations would build the corresponding target Dutch f-structure, which would be linked by the target language LFG-DOT model to the appropriate c-structure tree, as in Figure 4.

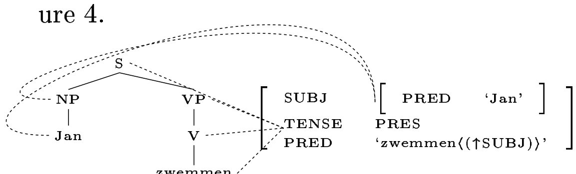


Figure 4: Robustness via *Discard* in LFG-DOT

The 'translation' of the ill-formed string *John swim* would therefore be *Jan zwemmen*. This shows that this particular DOT problem of boundary friction is resolved in LFG-DOT, owing to the presence of the syntactic f-structure information. We report further on the outstanding effect of boundary friction in LFG-DOT in section 4.

3.2 LFG-DOT Model 2: Translation via τ and γ

LFG-DOT2 requires the integration of the τ and γ mappings. Maintaining an f to f' translation engine in addition to γ increases the likelihood of achieving the correct translation—even if this is not proposed as the most probable translation via τ, given that this function will only ever produce very few translation candidates, we can guarantee in almost all cases that it is suggested as one of a small set of candidate translations. These can be compared to the best translation generated by γ and the highest ranking overall translation selected as output.

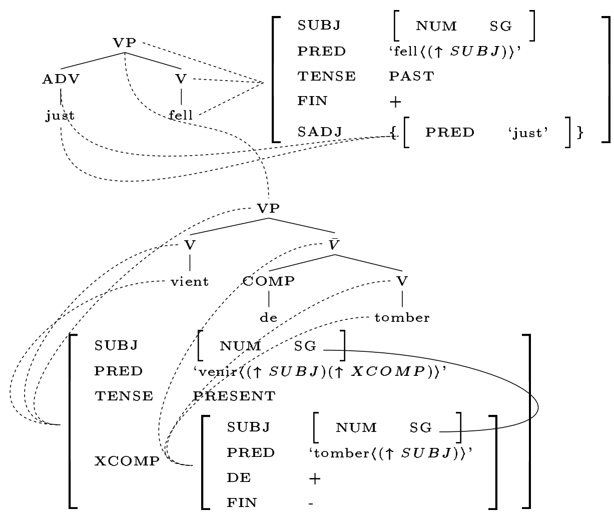


Figure 5: The *just* ⟷ *venir de* case in LFG-DOT3

However, the τ mapping cannot always produce the desired translation, so that most of the LFG-MT problems (notably, failure to deal successfully with certain headswitching examples, cf. Figure 5–Figure 7) are imported into both LFG-DOT1 and LFG-DOT2 models of translation.

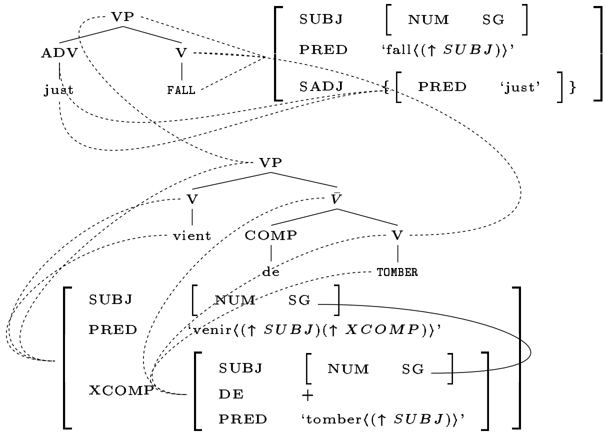


Figure 6: Lemmatization in LFG-DOT4

In contrast, Way (2001) shows that γ is always (depending, of course, on the coverage in the treebank) able to produce the correct translation, along with some possible wrong alternatives. The next two LFG-DOT models, therefore, abandon τ-equations and rely solely on γ to express the translation relation.

3.3 LFG-DOT Model 3: Translation via γ with Monolingual Filtering

The LFG-DOT3 model contains the DOT2 links between source and target c-structures, but with additional syntactic functional constraints which prevent ungrammatical structures such as Figure 3 from being formed (except via *Discard*, in much the same way as in (12) above), thereby enabling truly grammatical translations to be output, as opposed to translations which are grammatical only 'with respect to the corpus'. The f-structure information can be seen, therefore, as useful for monolingual disambiguation in both source and target sides. Ill-formed or unknown input is still processable by running *Discard* over the set of linked source and target $\langle c$, LFG-DOP-$\phi, f\rangle$ fragments.

As an example, consider the *just* $\longleftrightarrow$ *venir de* headswitching case. In terms of LFG-DOT3, the translation relation is shown in Figure 5.

The semantically equivalent source and target c-structures are linked at the VP level via γ (omitted here for reasons of clarity). We do not consider *fell* to be semantically equivalent to *tomber* owing to their different FIN(ite) values, added to the fact that *fell* has a TENSE value whilst *tomber* does not. Hence this translation fragment can only be reused by substituting this pair with associated singular NP subjects at the appropriate nodes in an S-linked fragment. In this respect, as with DOT2, this LFG-DOT3 model continues to suffer from limited compositionality. We address this concern further in the next section which describes the LFG-DOT4 model.

3.4 LFG-DOT Model 4: Translation via γ and 'Extended Transfer'

In the previous section, we observed that the outstanding problem with LFG-DOT3 is its retention of the DOT2 problem of limited compositionality. Returning to the *just* $\longleftrightarrow$ *venir de* headswitching case in Figure 5, we would like to be able to 'relax' some of the constraints in order to map $\langle fell, tomber\rangle$ to make these linked fragments more general, and hence more useful. In so doing, we would remove this problem of limited compositionality.

In LFG-DOT4, the basic translation relation is expressed by γ, as in LFG-DOT3. In LFG-DOT4, however, there is a second application of *Discard*, by which 'lemmatized' forms are arrived at on which 'extended transfer' can be performed. *Discard* relaxes constraints in order to produce a set of generalized fragments with the potential to deal with ill-formed or unknown input. Once the `TENSE` and `FIN` features have been relaxed on the lowest verbs in both fragments in Figure 5, they can be regarded as translationally equivalent. Given this, $\langle fell, tomber\rangle$ are linked and lemmatized, as in Figure 6.

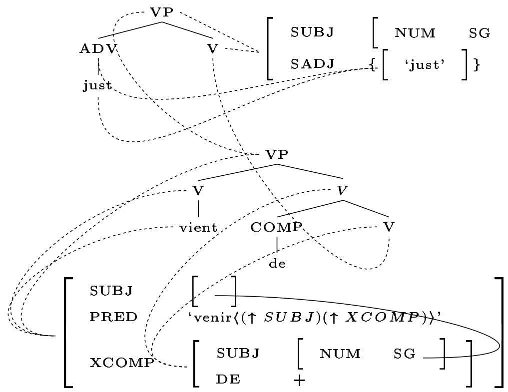


Figure 7: Generalized form of the *just* ⟷ *venir de* translation relation in LFG-DOT4

Now that ⟨FALL,TOMBER⟩ are linked, they can be deleted to produce the generalized form of the translation relation in Figure 7, as required. If fragment pairs such as those in Figure 7 prove subsequently to be of use in combining with other fragments, any resultant translation will be marked as ungrammatical with respect to the corpus, given that *Discard* was used in its derivation. Nevertheless, even if we restrict the impact of *Discard* on the probability space (cf. Way 2001, in order to ensure a preference for well-formed analyses derived via *Root* and *Frontier* over those produced by *Discard*), such translations will receive *some* probability, whereas the semi-compositional variants from which they were derived may not be able to produce *any* translation in practice.

4 Boundary Friction in LFG-DOT: two Experiments

Both DOT and LFG-DOT have strict definitions of fragment boundaries. We showed that whereas DOT cannot distinguish well-formed from ill-formed structures, LFG-DOT has an intuitively correct notion of grammaticality. Nevertheless, the thorny issue of boundary friction does raise its head in LFG-DOT to a degree.

All MT systems have to decide what are legitimate translation candidates. In most rule-based systems, default rules are differentiated from specific rules, with the former applying only in those cases where a specific rule cannot. Watanabe (1994) discusses the problem of boundary friction (he calls it 'example interference'), and provides a method of distinguishing exceptional from general examples in EBMT on the basis of *similarity* of examples (cf. Sato & Nagao, 1990, who use a similar technique based on thesaurus relations). Once patterns are identified as general, exceptional or neutral, some of the side-effects of boundary friction may be overcome.

Not all rule-based systems can prevent the output of a wrong, compositional translation once a specific translation has been obtained. For instance, in LFG-MT, satisfying the requirement that only possible translations are produced is problematic where the translation of a lexical head is conditioned in some way by one of its dependants, as in (13):

(13) commit suicide ⟷ se suicider

The problem is that in these cases, suppressing the wrong, compositional translation in LFG-MT is impossible. For instance, we require the default rules in (14):

(14) a. commit ⟷ commettre

b. suicide ⟷ suicide

Such rules are expressed in LFG-MT by the lexical entries in (15):

(15)

commit: $(\tau\uparrow \text{PRED}) = \text{commettre}$
$\tau(\uparrow \text{SUBJ}) = (\tau\uparrow \text{SUBJ})$
$\tau(\uparrow \text{OBJ}) = (\tau\uparrow \text{OBJ})$

suicide: $(\tau\uparrow \text{PRED}) = \text{suicide}$

These entries show how *commit* and *suicide* are to be translated under normal circumstances, such as in (16):

(16) a. Jean commet un crime $\longleftrightarrow$ John commits a crime

b. Le suicide est tragique $\longleftrightarrow$ Suicide is tragic

Nevertheless, given the default, compositional entries in (15), LFG-MT produces the wrong translation in (17):[2]

(17) John commits suicide $\longleftrightarrow$ *Jean commet le suicide

LFG-MT can, however, derive the correct translation *John se suicide* in such cases via the solution in (18):

$$\text{(18)} \qquad \textit{commit}: \quad \begin{array}{l} (\tau\!\uparrow \text{PRED}) = \text{se suicider} \\ \tau(\uparrow \text{SUBJ}) = (\tau\!\uparrow \text{SUBJ}) \\ (\uparrow \text{OBJ PRED}) =_c \text{suicide} \end{array}$$

Here the collocational units '*commit + suicide*' are linked as a whole to *se suicider*. The $=_c$ equation is a constraining equation: rather than expressing mere equality, it constrains the PRED value of the OBJ of *commit* to *suicide* when it is to be translated as a whole into *se suicider*. The selective use of constraining equations enables correct translations to be derived which would only be possible in other systems by tuning. Nevertheless, the point remains that in LFG-MT we would get both translations here, i.e. a correct one and a wrong one, since it is not possible to enforce the requirement that specific rules ought to override the default translational rules where applicable.

4.1 Experiment 1

We tested the issue of default versus specific translations in LFG-DOT3. We produced an LFG-DOT3 treebank containing all the linked fragments from the sentences in (19):

[2]Note that the rules in (14) are *bona fide* translation rules that any rule-based English-French MT system will require. It is, therefore, the task of the French generation component to explicitly rule out the incorrect translation in (17), *not* the transfer component.

(19) a. Le suicide est tragique $\longleftrightarrow$ Suicide is tragic.

b. Jean commet le crime $\longleftrightarrow$ John commits the crime.

c. Jean commet le meurtre $\longleftrightarrow$ John commits the murder.

d. Jean dort $\longleftrightarrow$ John sleeps.

e. Marie se suicide $\longleftrightarrow$ Mary commits suicide.

f. Marie commet un attentat $\longleftrightarrow$ Mary commits an attack.

g. Marie commet la faute $\longleftrightarrow$ Mary commits the mistake.

h. Pierre commet un arbitre $\longleftrightarrow$ Peter nominates an arbitrator.

i. Pierre commet une erreur $\longleftrightarrow$ Peter commits an error.

j. Pierre commet une injustice $\longleftrightarrow$ Peter commits an injustice.

Here there are seven instances of *commettre* (six of which translate as *commit*) as opposed to only one instance of *se suicider*.

Before examining results obtained with LFG-DOT, it is insightful to point out that in the monolingual French LFG-DOP (no *Discard*) treebank built from the French strings in (19), *Marie se suicide* is preferred about 2.6 times over the compositional alternative *Marie commet le suicide*. If these were the output translations, then ranking them against one another would favour *Marie se suicide* with about 72% versus *Mary commet le suicide* with about 28%. In the LFG-DOT3 treebank produced from the English and French sentences in (19), the specific translation is preferred even more than in the French monolingual LFG-DOP treebank. We set out to test the weight of the specific over the compositional translation for the sentences in (20):

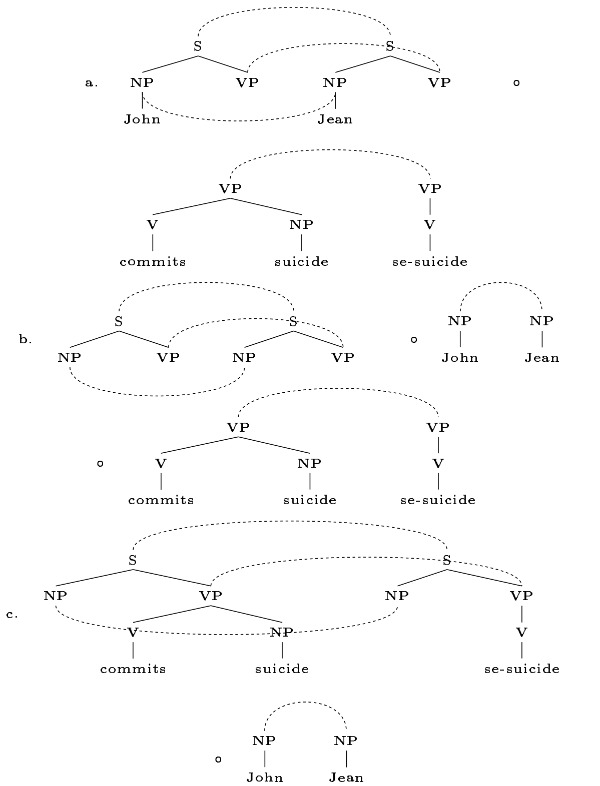


Figure 8: C-structure derivations for *John commits suicide* $\Longleftrightarrow$ *Jean se suicide*

(20) a. John commits suicide $\longleftrightarrow$ Jean se suicide

b. Mary commits suicide $\longleftrightarrow$ Marie se suicide

c. John commits suicide $\longleftrightarrow$ *Jean commet le suicide

d. Mary commits suicide $\longleftrightarrow$ *Marie commet le suicide

Translation (20a) can be built using the three derivations in Figure 8.[3] (20b) has the additional derivation of the full trees (and accompanying f-structures) for this sentence pair. The probabilities of (20a-b) are shown in (21):

(21) a. P(John commits suicide $\longleftrightarrow$ Jean se suicide) $= 0.000705$ ($\simeq \frac{1}{1419}$)

b. P(Mary commits suicide $\longleftrightarrow$ Marie se suicide $= 0.006229$ ($\simeq \frac{1}{161}$)

[3]We have omitted the accompanying f-structure fragments for reasons of space.

For each of the translations in (20c-d) there are 7 derivations with total probability 0.000501 ($\simeq \frac{1}{1998}$). Now we can rank each translation with respect to the other in (22):

(22) a. P(John commits suicide $\stackrel{T}{=}$ Jean se suicide) $= 705/1206 = 0.5846$

b. P(John commits suicide $\stackrel{T}{=}$ Jean commet le suicide) $= 501/1206 = 0.4154$

c. P(Mary commits suicide $\stackrel{T}{=}$ Marie se suicide $= 6229/6750 = 0.923$

d. P(Mary commits suicide $\stackrel{T}{=}$ Marie commet le suicide) $= 521/6750 = 0.077$

Here $\mathcal{T} \stackrel{T}{=} W$ means that $\mathcal{T}$ is a translation of a word string W. Therefore we can see that for *John commits suicide*, the correct, specific translation is about 1.4 times more likely than the wrong, default, compositional translation, whereas for *Mary commits suicide* the specific translation is preferred about 12 times more than the default translation. We see in (22) the dominance of the exact linked translation pair over the alternative translation. The presence of the exact translation (19e) is insufficient to explain the preference for the specific translation for *John commits suicide*: despite the presence of six *commit* $\longleftrightarrow$ *commettre* examples in (19) compared to only the single instance of *commits suicide* $\longleftrightarrow$ *se suicide*, the specific translation is nonetheless preferred.

4.2 Experiment 2

How many more times empirically we can expect to see *commit* $\longleftrightarrow$ *commettre* compared to *commits suicide* $\longleftrightarrow$ *se suicide*? In the LOB Corpus, there are 66 instances of *commit* (including its morphological variants), only 4 of which have *suicide* as its object, out of the 15 occurrences of *suicide* as an NP. Consequently, even for this small sample, we can see that 94% of these examples need to be translated compositionally (by *commettre* + NP), while only the *commit suicide* examples require a specific rule to apply (i.e. *se suicider*).

In the on-line Canadian Hansards covering 1986-1993, there are just 106 instances of *se suicider* (including its morphological variants). There will, of course, be many more instances of *commettre*. Given occurrences of *suicide* as an NP in French corpora, it is not an unreasonable hypothesis to expect that wrong translations such as (17) will be much more probable than those derived via the specific rule. However, this hypothesis is shown to be inaccurate in the above experiment.

Furthermore, it is clear from the results in (22) that a ratio of 6:1 is insufficient to achieve a bias in favour of the wrong, compositional translation in LFG-DOT. Running a new experiment with a treebank built from 5 instances of each translation pair (19a-d) and (19f-j) and just the one instance of (19e), making a total of 46 sentences in all, produces the results in (23):

(23) a. P(John commits suicide $\stackrel{T}{=}$ Jean se suicide) = $132/635 = 0.208$

b. P(John commits suicide $\stackrel{T}{=}$ Jean commet le suicide) = $503/635 = 0.792$

c. P(Mary commits suicide $\stackrel{T}{=}$ Marie se suicide = $1206/1758 = 0.686$

d. P(Mary commits suicide $\stackrel{T}{=}$ Marie commet le suicide) = $552/1758 = 0.314$

Now, with 30 instances of *commit* $\longleftrightarrow$ *commettre* and only the one *commits suicide* $\longleftrightarrow$ *se suicide* example, we see that the wrong, default, compositional translation for *John commits suicide* is now preferred by about 3.8 times, but the presence of the exact translation (19e) maintains the preference for the specific translation for *Mary commits suicide* by about 2.2 times. Consequently we can see that it will take many more instances of *commit* $\longleftrightarrow$ *commettre* before the specific translation for *Mary commits suicide* is outranked by the wrong, compositional alternative.

5 Conclusions and Future Work

Models of translation based on DOP and LFG-DOP translate new strings on the basis of linked ⟨*source, target*⟩ fragments already located in their databases. Accordingly, such systems may be viewed as example-based systems.

We described the DOT models of translation based on DOP. DOT1 is not guaranteed to produce the correct translation when this is non-compositional and considerably less probable than the default, compositional alternative. DOT2 addresses the failings of DOT1 by redefining the composition operation. In contrast to DOT1, DOT2 cannot fail to produce correct candidate translations, along with some possible wrong alternatives, depending of course on the corpus from which fragments are derived. Despite the presence of syntactic information in the tree-structure fragments, we showed that both DOT models continue to suffer from the problem of boundary friction in cases where singular and plural fragments are combined.

We also described a number of new hybrid models of translation which use LFG-DOP as their language models. The first, LFG-DOT1, imports the τ-equations from LFG-MT as the translation relation. LFG-DOT1 improves the robustness of LFG-MT through the use of the LFG-DOP *Discard* operator, which produces generalized fragments by discarding certain f-structure features. It can, therefore, deal with ill-formed or previously unseen input where LFG-MT cannot. Unsurprisingly, however, all of the other problems of LFG-MT are maintained in LFG-DOT1.

Given this, we augmented LFG-DOT1 with the γ function from DOT2 to give an improved model of translation. LFG-DOT2 maintains the τ translation relation to increase the chances of the correct translation being produced. Nevertheless, given that the τ-equations fail to derive the correct translation in all cases, we omitted the τ translation relation from our subsequent models.

LFG-DOT3 relies wholly on γ to express the translation relation, and uses f-structure information purely for monolingual filtering. The presence of this functional information prevents the formation of certain ill-formed structures which can be produced in DOT. LFG-DOT models, therefore, have a notion of grammaticality which is missing from DOT models. While both DOT and LFG-DOT contain strict notions of boundary definition, DOT allows the output of structures which are well-formed according to the corpus, but which are syntactically ungrammatical. The definition of well-formedness in LFG-DOT, in contrast, corresponds exactly to our understanding of grammaticality in the wider sense. However, both DOT2 and LFG-DOT3 models suffer from limited compositionality, so that in some cases the minimal statement of the translation relation is impossible.

LFG-DOT4 adds an 'Extended Transfer' phase to LFG-DOT3 by producing lemmatized forms using a second application of *Discard*. This extension overcomes the problem of limited compositionality, enabling the statement of the translation relation in an intuitive, concise fashion.

Finally, we demonstrated that LFG-DOT models of translation suffer less from the problem of boundary friction than DOT models given the presence of the additional syntactic f-structure information. In addition, we showed in two small experiments that despite attempting to 'load the dice' in favour of the wrong, compositional translation over the correct, specific alternative, LFG-DOT continues to translate in a robust fashion.

The work described here and in Way (2001) uses as its evaluation metric the ability to cope with 'hard' translation cases, such as relation-changing (cf. (5)–(6)) and headswitching (cf. (7)–(10)). Special LFG-DOT corpora such as those derived to test the effect of Boundary Friction in section 4 needed to be created. The translation effects examined here need to be tested further on larger corpora, and the work of Frank *et al.* (2001) on semi-automatic derivation of LFG corpora from treebank resources would appear promising in this regard.

Furthermore, the hypotheses developed in Way (2001) need to be further explored with respect to simpler translation data, such as *Fido barks* sentences. Different probability models will also be evaluated (cf. Bonnema *et al.*, 2000), as will the possibility of pruning the search space, by cutting down the number of fragments produced (cf. Sima'an, 1999) in order to improve the efficiency of the models proposed.

References

[1] Bod, R. (1998): *Beyond Grammar: An Experience-Based Theory of Language*, CSLI, Stanford, CA.

[2] Bod, R. and R. Kaplan (1998): 'A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis', in *Proceedings of the 17th International Conference on COLING and 36th Conference of the ACL*, Montreal, Canada, **1**:145–151.

[3] Bonnema, R., P. Buying and R. Scha (2000): 'Parse Tree Probability in Data-Oriented Parsing', in *Proceedings of PACLING*, Mexico City.

[4] Frank, A., J. van Genabith and A. Way (2001): 'Treebank vs. X-BAR based Automatic F-Structure Annotation', in *Proceedings of 6th International Conference on Lexical Functional Grammar (LFG-2001)*, Hong Kong.

[5] Kaplan, R. and J. Bresnan, (1982): 'Lexical Functional Grammar: A Formal System for Grammatical for Grammatical Representation', in J. Bresnan (ed.) *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Mass., pp.173–281.

[6] Kaplan, R., K.. Netter, J. Wedekind and A. Zaenen (1989): 'Translation by Structural Correspondences', in *Fourth Conference of the EACL*, Manchester, pp.272–281.

[7] Mima, H, H. Iida and O. Furuse (1998): 'Simultaneous Interpretation Utilizing Example-Based Incremental Transfer', in *Proceedings of the 17th International Conference on COLING and 36th Conference of the ACL*, Montreal, Canada, pp.855-861.

[8] Poutsma, A. (1998): 'Data-Oriented Translation', in *Ninth Conference of Computational Linguistics In the Netherlands*, Leuven, Belgium.

[9] Poutsma, A. (2000): *Data-Oriented Translation: Using the Data-Oriented Parsing framework for Machine Translation*, MSc Thesis, University of Amsterdam, The Netherlands.

[10] Sato, S. and M. Nagao (1990): 'Towards Memory-based Translation', in *13th International Conference on COLING*, Helsinki, Finland, **3**:247–252.

[11] Sima'an, K. (1999): *Learning Efficient Disambiguation*, PhD Thesis, University of Utrecht, The Netherlands.

[12] Somers, H. (1999): 'Example-based Machine Translation', *Machine Translation* **14**(2):113–157.

[13] Somers, H., I. McLean and D. Jones (1994): 'Experiments in Multilingual Example-Based Generation', in *Proceedings of the 3rd Conference on CSNLP*, Dublin, Ireland.

[14] Somers, H., J. Tsujii and D. Jones (1990): 'Machine Translation without a source text', in *13th International Conference on COLING*, Helsinki, Finland, **3**:271–276.

[15] Watanabe, H. (1992): 'A Similarity-Driven Transfer System', in *14th International Conference on COLING*, Nantes, Frances, pp.770–776.

[16] Watanabe, H. (1994): 'A Method for Distinguishing Exceptional and General Examples in Example-based Transfer Systems', in *15th International Conference on COLING*, Kyoto, Japan, pp. 39–44.

[17] Way, A. (2001): *LFG-DOT: A Hybrid Architecture for Robust MT*, PhD Thesis, University of Essex, UK.