May 1991

Using Lexicalized Tags for Machine Translation

Anne Abeillé
University of Paris

Yves Schabes
University of Pennsylvania

Aravind K. Joshi
University of Pennsylvania, joshi@cis.upenn.edu

University of Pennsylvania Department of Computer and Information Sciences Technical Report No. MS-CIS-91-44.

Using Lexicalized Tags for Machine Translation

Abstract

Lexicalized Tree Adjoining Grammar (LTAG) is an attractive formalism for linguistic description mainly because of its extended domain of locality and its factoring recursion out from the domain of local dependencies (Joshi, 1984, Kroch and Joshi, 1985, Abeillé, 1988). LTAG's extended domain of locality enables one to localize syntactic dependencies (such as filler-gap), as well as semantic dependencies (such as predicate-arguments). The aim of this paper is to show that these properties combined with the lexicalized property of LTAG are especially attractive for machine translation. The transfer between two languages, such as French and English, can be done by putting directly into correspondence large elementary universe without going through some interlingual representation and without major changes to the source and target grammars. The underlying formalism from the transfer is "synchronous Tree Adjoining Grammars" (Sheiber and Schabes [1990]). Transfer rules are stated as correspondences between nodes of trees of large domain of locality which are associated with words. We can thus define lexical transfer rules that avoid the defects of a mere word-to-word approach but still benefit from the simplicity and elegance of a lexical approach. We rely on the French and English LTAG grammars (Abeillé [1988], Abeillé [1990(b)], Abeillé et al. [1990], Abeillé and Schabes [1989, 1990]) that have been designed over the past two years jointly at University of Pennsylvania and University of Paris 7-Jussieu.

Comments

University of Pennsylvania Department of Computer and Information Sciences Technical Report No. MS-CIS-91-44.

Using Lexicalized Tags For Machine Translation

MS-CIS-91-44
LINC LAB 204

Anne Abeillé
(University of Paris 7-Jussieu)

Yves Schabes
(University of Pennsylvania)

Aravind K. Joshi
(University of Pennsylvania)

Department of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104-6389

May 1991

Proceedings of the *International Conference on Computations Linguistics* (COLING - 90), Helsinki, Finland, August 1990

Using Lexicalized Tags for Machine Translation *

Anne Abeillé
University of Paris 7-Jussieu
LADL
2 place Jussieu, 75005 Paris France
abeille@franz.ibp.fr

Yves Schabes
University of Pennsylvania
Dept of Computer & Information Science
Philadelphia PA 19104-6389
schabes@linc.cis.upenn.edu

Aravind K. Joshi
University of Pennsylvania
Dept of Computer & Information Science
Philadelphia PA 19104-6389
joshi@linc.cis.upenn.edu

Abstract

Lexicalized Tree Adjoining Grammar (LTAG) is an attractive formalism for linguistic description mainly because of its extended domain of locality and its factoring recursion out from the domain of local dependencies (Joshi, 1985, Kroch and Joshi, 1985, Abeillé, 1988). LTAG's extended domain of locality enables one to localize syntactic dependencies (such as filler-gap), as well as semantic dependencies (such as predicate-arguments). The aim of this paper is to show that these properties combined with the lexicalized property of LTAG are especially attractive for machine translation.

The transfer between two languages, such as French and English, can be done by putting directly into correspondence large elementary units without going through some interlingual representation and without major changes to the source and target grammars. The underlying formalism for the transfer is "synchronous Tree Adjoining Grammars" (Shieber and Schabes [1990])[1]. Transfer rules are stated as correspondences between nodes of trees of large domain of locality which are associated with words. We can thus define lexical transfer rules that avoid the defects of a mere word-to-word approach but still benefit from the simplicity and elegance of a lexical approach.

We rely on the French and English LTAG grammars (Abeille [1988], Abeille [1990 (b)], Abeillé et al. [1990], Abeillé and Schabes [1989, 1990]) that have been designed over the past two years jointly at University of Pennsylvania and University of Paris 7-Jussieu.

1 Strategy for Machine Translation with LTAGs

The idea of using grammars written with "lexicalist" formalisms for machine translation is not new and has been exemplified by Kaplan, et al., (1989) for LFG, Beaven et al. for UCG (1988), Dorr for GB (1989) and Arnold et al. for Eurotra (1986). However, our approach is more radical in the sense that we associate with the lexical items structures that localize syntactic and semantic dependencies. This allows for the possibility that an explicit semantic representation level can be avoided.[2] The claims about the advantages of an explicit semantic representation level need to be investigated again in the context of the approach proposed here. For examples, many traditionally difficult problems for machine translation due to different divergence types (Dorr 1989) such as categorial, thematic, conflational, structural and lexical are not problems in the approach we suggest. Also contrary to UCG, but like LFG, we use grammars that have not been designed for the purpose of translation.

The underlying formalism achieving the transfer of derivations is "Synchronous Tree-Adjoining Grammars" (as described in a companion paper by Shieber and Schabes [1990]).[3] The strategy adopted for machine translation consists of matching the source LTAG derivation of the source sentence to a target LTAG derivation by looking at a transfer lexicon. The transfer lexicon puts into correspondence a tree from the source grammar instantiated by lexical insertion (all its nodes and their attributes) with a tree from the target grammar. Although the approach is not inherently directional, for convenience we will call the English and French grammars, the source and target grammars.

The translation process consists of three steps in which the generation step is reduced to a trivial step. First the source sentence is parsed accordingly to the source grammar. Each elementary tree in the derivation is now considered with the features given from the derivation through unification. Second, the source derivation tree is transferred to a

*This research was partially funded by ARO grant DAAG29-84-K-0061, DARPA grant N00014-85-K0018, and NSF grant MCS-82-19196 at the University of Pen nsylvania. We are indebted to Stuart Shieber for his valuable comments. We would like also to thank Marilyn Walker.

[1] In this volume.

[2] The formalism of Synchronous Tree-Adjoining Grammar does not prevent constructing an explicit semantic representation. In fact, in Shieber and Schabes (1990) it is shown how to construct a semantic representation, which itself is a TAG.

[3] We assume that the reader is familiar with Tree Adjoining Grammars. We refer the reader to Joshi (1987) for an introduction to TAGs. We also refer the reader to the companion paper for more details on synchronous TAGs.

target derivation. This step maps each elementary tree in the source derivation tree to a tree in the target derivation tree by looking in the transfer lexicon. And finally, the target sentence is generated from the target derivation tree obtained in the previous step.

As an example, consider the fragment of the transfer lexicon given in Figure 1.

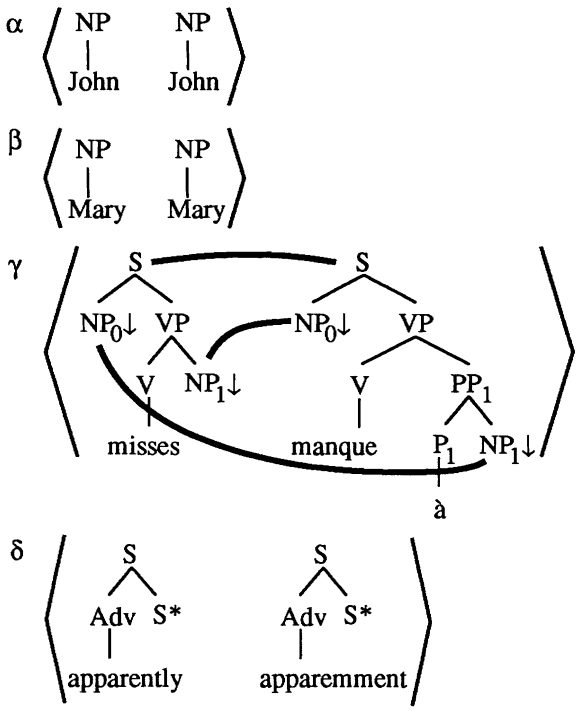
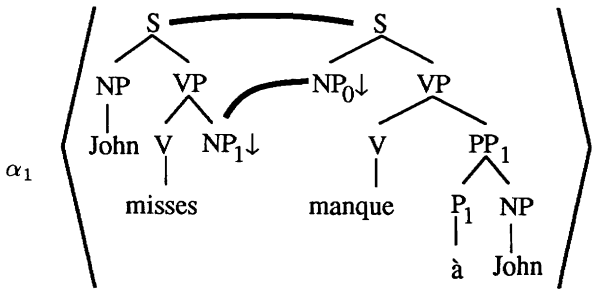
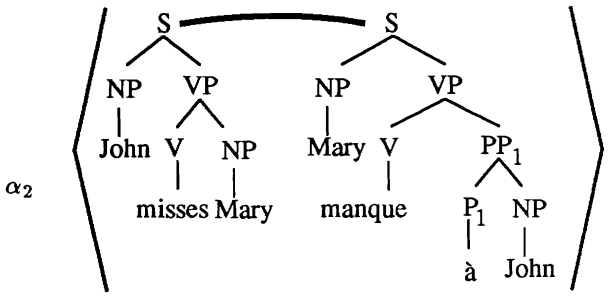


Figure 1: Fragment of the English-French transfer lexicon

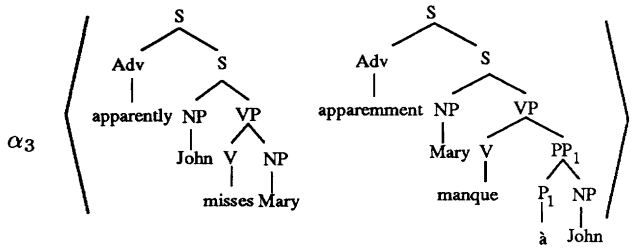
The transfer lexicon consists of pairs of trees one from the source language and one from the target language. Within the pair of trees, nodes may be linked (thick lines). Whenever in a source tree, say t_{source}, adjunction or substitution is performed on a linked node (say n_{source} is linked to n_{target}) , the corresponding tree paired with t_{source}, t_{target}, operates on the linked node n_{target}. For example, suppose we start with the pair γ and we operate the pair α on the link from the English node NP_0 to the French node NP_1. This operation yields the derived pair α_1.



Then, if the pair β operates on the NP_1-NP_0 in α_1, the following pair α_2 is generated.



Finally, when the pair δ operates on the S-S link in α_2, the pair α_3 is generated.



The source sentence is parsed accordingly to the source grammar, then the target derivation is generated by tracing the pairs stated in the transfer lexicon. The fragment of the transfer lexicon given in Figure 1 therefore enables us to translate:

Apparently, John misses Mary
↔ *Apparemment, Mary manque à John*

In most cases, translation can be performed incrementally as the input string is being parsed.

The aim of this paper is to show that LTAG's localization of syntactic dependencies (such as filler-gap), as well as semantic dependencies (such as predicate-arguments) combined with the lexicalized property of LTAGs are especially attractive for machine translation.

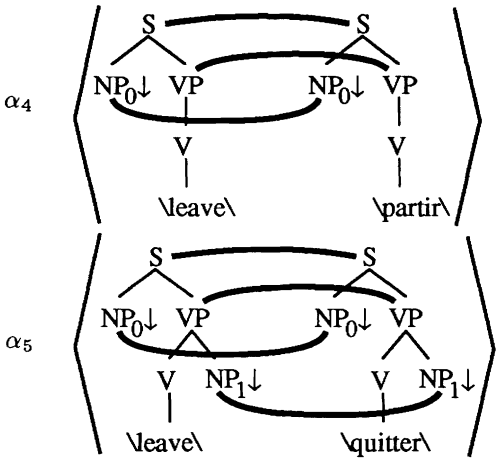
We show how the transfer lexicon is stated. We motivate the need for mapping trees instantiated with words and with the value of their features obtained from the derivation tree corresponding to the parse of the source sentence. We also show that the transfer needs to be stated at different levels: matching tree families (trees associated to the same predicate), trees, nodes and therefore their attributes, since they are associated with a node. We show how not only subcategorization frames but also adjuncts are transferred, and how differences of syntactic and semantic properties are accounted for in terms of structural discrepancies. Then we illustrate how the extended domain of locality enables us to deal with these structural discrepancies in the process of machine translation.

2 Transfer Lexicon– matching two LTAG Lexicons

The transfer is stated between the English and French LTAG grammars in a lexicon. We rely on grammars built from a monolingual perspective, but the match between them can be one to many, or many to one.

2.1 Matching elementary trees

Instead of matching words, we match structures in which words have been already lexically inserted. This provides interesting disambiguations that could not be obtained by a morphological match. For example, there is one morphological English verb *leave*, but the structures associated with it disambiguate it between intransitive and transitive *leave*. Interestingly, these two predicates receive two different French translations: [4]



The pairs α_4 and α_5 will correctly give the following translations:

John left ↔ *John est parti*

John left Mary ↔ *John a quitté Mary*

By convention, in the elementary trees, the set of morphological flexions of a given word is written surrounded by backslashes. For example, \leave\ stands for {*leave, leaves, left,* ···}. For each word in a morphological set attributes (such as mode and agreement) are also specified. When a word in a tree is not surrounded by backslahes, it stands for the inflected form and not for a morphological set.

Since lexical items appearing in the elementary structures can be inflected words or a morphological set, lexical items of the two languages are matched regardless of whether they exhibit the same morphological variations or not. For example, English adjectives lacking morphological variation appear as such in the syntactic and transfer lexicons, while their French counterparts are usually morphological sets. The word *white* is thus matched with \blanc\, standing for {*blanc, blanche, blancs, blanches*}.

Words that are not autonomous entries in the English syntactic lexicon (ex: complementizers, light verbs or parts of an idiomatic expression), are not considered as autonomous entries in the transfer lexicon; for example, no rule needs to match directly *take* or *pay* with *faire*, or *give* with *pousser*, in order to get the right light-verb predicative noun combinations in the following sentences:[5]

John took a walk
↔ *John a fait une promenade (Danlos 1989)*

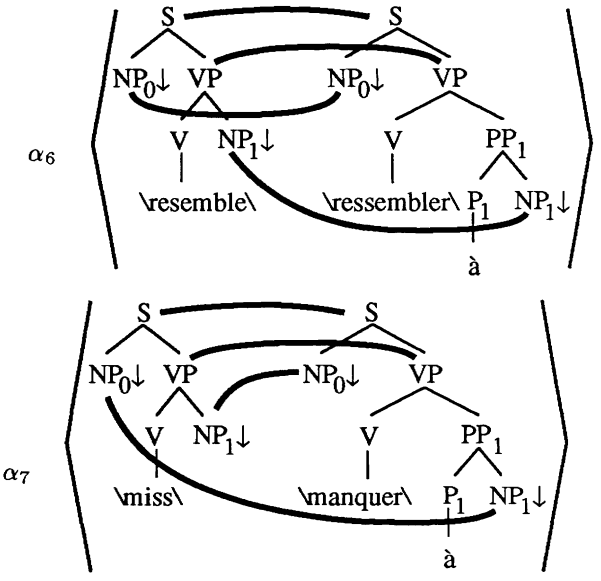
John pays court to Mary
↔ *John fait la court à Mary (Danlos 1989)*

John gave a cry ↔ *Jean a poussé un cri*

Some words existing as autonomous entries in the English syntactic lexicon do not appear as entries in the transfer lexicon because their French counterpart is a morphological flexion, not a word. For example, the future auxiliaries *will* or *shall* are not translated as such. The tense feature they contribute is transferred (as well other syntactic features) and the future tense French verbal form will be chosen.

2.2 Matching nodes

Matching predicates of the two languages as a whole is not sufficient. Correspondences between their arguments must be stated too as shown in the following example:



John resembles Mary ↔ *John ressemble à Mary*

John misses Mary ↔ *Mary manque à John*

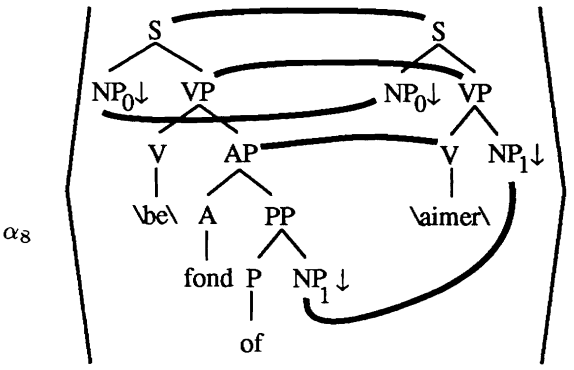
[4] We use standard TAG notation: '↓' stands for nodes to be substituted, '*' annotates the foot node of an auxiliary tree and the indices shown on the nodes correspond to semantic functions. The trees are combined with adjunction and substitution.

Our approach does not depend on the specific representation adopted in this paper. See Abeillé 1990 (b) for an alternate representation.

[5] It has long been noticed that light-verb predicative noun combinations are highly language-idiosyncratic, and word-to-word transfer rules will inevitably lead to overgeneration or unnatural restrictions.

These examples also show that it is not correct to match trees where lexical insertion has not already been made and therefore the correspondences between nodes cannot be made on the only basis of the subcategorization frame.

Arguments are matched directly by the links existing between them. Adjuncts are matched indirectly by the links existing on the nodes, at which they adjoin. For example, in the following correspondence,



the AP node in the English tree is linked to the V node of the French tree to account for:

John is fond of music
↔ *John aime la musique*

John is very fond of music
↔ *John aime beaucoup la musique*

The adjective *fond* is associated with an AP-type auxiliary tree which is paired with a V-type auxiliary tree corresponding to the word *beaucoup*.

2.3 Matching feature structures

Some feature structures of the words appearing in the trees are transferred in the translation process, but with the value further specified from the derivation (and not with the one from the lexical entry which may not be as specific). For example, *fish* can be either singular or plural and is therefore stated as such in the lexicon. However, it can get its number from the verb-subject agreement constraints, as in the following sentences:

The fish swim in the pond
↔ *Les poissons nagent dans l'étang (plural)*

The fish is good
↔ *Le poisson est bon (singular)*

Agreement features of nouns are lexically matched only in the case of two morphological sets. In the case of one (or both) entry being a single inflected word, the agreement features depend only on the lexical entry itself and are directly assigned in the transfer lexicon:

\boy\,N [num=X] ↔ *\garçon\,N [num= X]*
luggage, N [num=sing] ↔ *bagages,N [num = pl]*

Because of these idiosyncrasies, agreement features of verbs are not matched. We will thus rightly have:

My luggage is heavy (singular)
↔ *Mes bagages sont lourds (plural)*

based on monolingual agreement constraints between subject and verb.

Features assigned to the sentential root node (either from lexical insertion or from some adjoined material) are transferred or not depending on whether they are assigned autonomously in the target language or not. The feature *tense* for example is usually transferred, but not the feature *mode*, because the latter depends on the verb of the matrix sentence if the sentence is embedded:

Jean wants Marie to leave
↔ *Jean veut que Marie parte (Danlos 1989)*

2.4 Matching tree families

In order to transfer both the predicate-argument relations, and the construction types such as question, passive, topicalization etc., it is necessary to be able to refer to a specific tree in a tree family. This is done by matching the syntactic features by which the different trees are identified within a tree family, for example <passive>, <relative, NP_i > or <question, NP_i >.[6]

As has been noted, transitivity alternations exhibit striking differences in the two languages. The trees in the two families will not necessarily bear the same syntactic features; corresponding tree families may not include the same number of trees.

When a syntactic feature of a given tree family does not exist for the corresponding tree family in the target language, it will be ignored. English trees for prepositional passives will thus be matched with their corresponding declarative trees in French (unless the English prepositional argument is matched with the French direct object):

John was given a book by Mary
↔ *Mary a donné un livre à Jean*

Similarly, the feature <question, NP_i > will be transferred but not the feature differentiating between pied-piping and preposition-stranding in English, since French always pied-pipes:

Who did Mary give a book to?
↔ *A qui Mary a-t-elle donné un livre?*

When a certain syntactic feature exists for both tree families in the two languages, but not for both lexical items, it is ignored as well:

Advantage was taken of this affair by John
↔ * *Parti a été tiré de cette affaire par Jean*
↔ *Jean a tiré parti de cette affaire*

Such idiosyncrasies are in fact expected and handled in our grammars, since they have both their constituent structures and their syntactic rules lexicalized (see Abeillé [1990 (a)] for a discussion on this topic).

[6] NP_i refers to the noun phrase being extracted, usually 0 for subject, 1 for first object etc. .

3 Dealing with Structural Discrepancies

Units of a LTAG grammar have a large domain of locality. Discrepancies in the internal structures being matched are in fact expected by our strategy, and no special mechanism is required for them.

3.1 Discrepancies in constituent structures

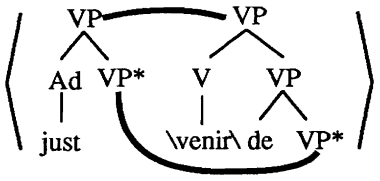
It is not a problem when an elementary tree of a certain constituent structure translates into an elementary tree with a different constituent structure in the target language, provided they have a similar argument structure. For example: idiom ↔ verb; idiom ↔ different kind of idiom; verb ↔ light-verb combination; VP-adverb ↔ raising verb; S-adverb ⇔ matrix clause ... as in:

The baby just fell
↔ *Le bébé vient de tomber (Kaplan et al. 1989)*

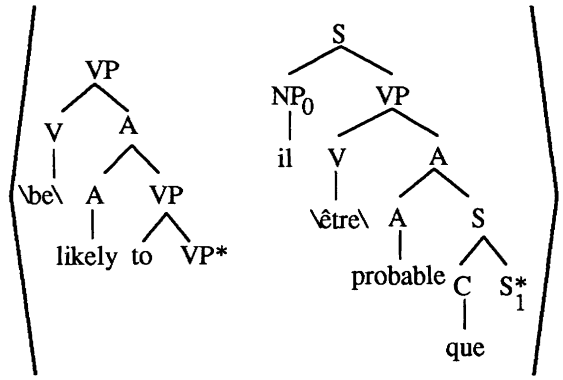
John is likely to come
↔ *Il est probable que Jean viendra*

John gave a cough ↔ *John toussa*

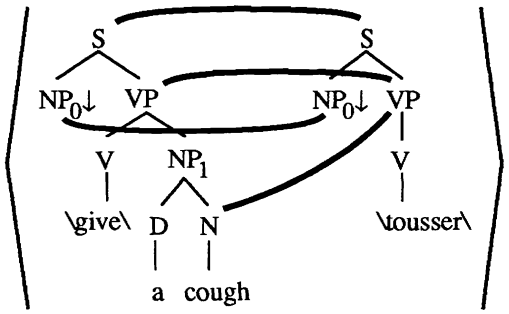
α_9



α_{10}



α_{11}



Links provide for simultaneous adjunction (or substitution) of matching trees at the corresponding nodes. For example in the pair α_{11}, adjunction of an adjective (on N) in the English tree corresponds to an adjunction on the French VP:

John gave a weak cough
↔ *John toussa faiblement*

Furthermore elementary structures of the source language need not exist in the target language as elementary structures. For example, there is no French counterpart to the English verb particle combination.

John called Mary up ↔ *John a appelé Mary*

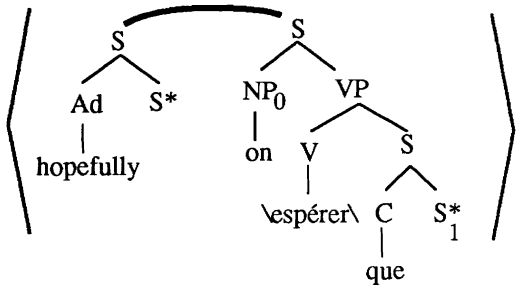
3.2 Discrepancies in syntactic properties

Some English predicates do not have the same number of arguments as their corresponding French ones. In such cases, the pair does not consists of pairs of elementary trees but rather pairs of derived trees of bounded size. Since the match is performed between derived trees, no new elementary trees are introduced in the grammars. This addition of pairs of bounded derived trees is the only change we have to make to the units of the original grammars.

For example, the adverb *hopefully* has an S argument. Since there is no corresponding French adverb, the French verb *espérer* (which has two arguments, an NP and an S) combined with *on* will be used:

hopefully, John will work
↔ *on espère que Jean travaillera*

α_{12}



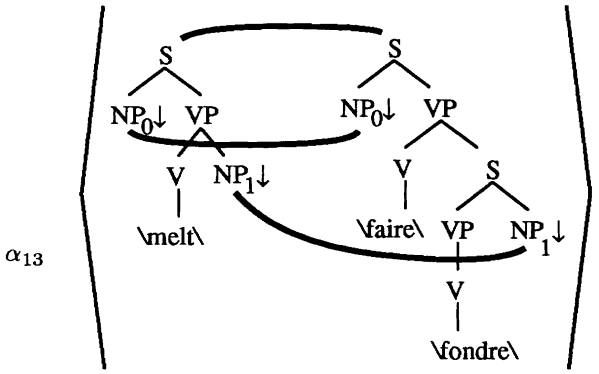
In the pair α_{12}, *hopefully* is paired with a derived tree corresponding to *on espère*. The English tree for *hopefully* is paired with the result of the substitution of *on* in the subject position of the tree for *espèrer*. The right hand tree in α_{12} is a derived tree.

Matching agentless passive with declarative trees is done with the same device:

John was given a book
↔ *On a donné un livre à John*

Similar cases occur for verbs exhibiting ergativity alternation in one language and but not in the other. In this case, a supplementary causative tree has to be used for the unaccusative language (see pair α_{13}):

The sun melts the snow
↔ * *le soleil fond la neige*
↔ *le soleil fait fondre la neige*



The right hand tree in α_{13} is again a derived tree.

Multicomponent TAG (Joshi [1987]) can also be used for resolving certain other discrepancies. This device is not a new addition, it is already a part of the Synchronous TAG framework.

Conclusion

By virtue of their extended domain of locality, Tree Adjoining Grammars allow regular correspondences between larger structures to be stated without a mediating interlingual representation. The mapping of derivation trees from source to target languages, using the formalism of synchronous TAGs, makes possible to state such direct correspondences. By doing so, we are able to match linguistic units with quite different internal structures. Furthermore, the fact that the grammars are lexicalized enables capturing some idiosyncrasies of each language.

The simplicity and effectiveness of the transfer rules in this approach shows that lexicalized TAGs, with their extended domain of locality, are very well adapted to machine translation. A detailed discussion of this approach will be provided in an expanded version of this paper which will include a discussion of the applicability of this method for other pairs of languages exhibiting some language phenomena that do not arise in the pair considered in this paper.

References

Abeillé, Anne. 1988. Parsing French with Tree Adjoining Grammar: some Linguistic Accounts. In *Proceedings of the* 12^{th} *International Conference on Computational Linguistics (COLING'88).* Budapest, Hungary.

Abeillé, Anne. 1990 (a). Lexical and Syntactic Rules in a Tree Adjoining Grammar. In *Proccedings of the* 28^{th} *Meeting of the Association for Computational Linguistics (ACL'90).* Pittsburgh, PA.

Abeillé, Anne. 1990 (b). A Lexicalized Tree Adjoining Grammar and its Relevance for Machine Translation. To appear in *Machine Translation.*

Abeillé, Anne and Schabes, Yves. 1989. Parsing Idioms in Tree Adjoining Grammars. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL'89).* Manchester.

Abeillé, Anne, Bishop, Kathleen M., Cote, Sharon, and Schabes, Yves. 1990. *A Lexicalized Tree Adjoining Grammar for English.* Technical Report, Department of Computer and Information Science, University of Pennsylvania.

Abeillé, Anne and Schabes, Yves. 1990. Non Compositional Discontinuous Constituents in Tree Adjoining Grammar In *Proceedings of the Symposium on Discontinuous.* Tilburg, Holland.

Arnold, D., Krauwer, S., Rosner, M., Destombes, L. and Varile, G. 1986. The CAT Framework in Eurotra: a Theoretically Committed Notation for Machine Translation. In *Proceedings of the* 11^{th} *International Conference on Computational Linguistics (COLING'86).* Bonn.

Beaven, John and Whitelock, Pete. 1988. Machine Translation Using Isomorphic UCGs. In *Proceedings of the* 12^{th} *International Conference on Computational Linguistics (COLING'88).* Budapest, Hungary.

Danlos, Laurence. 1989. La traduction automatique. *Annales des Télécommunications* 44(1-2).

Dorr, Bonnie J. 1989. Conceptual Basis of the Lexicon in Machine Translation. MIT AI lab Memo No 1166.

Joshi, Aravind K. 1985. How Much Context-Sensitivity is Necessary for Characterizing Structural Descriptions—Tree Adjoining Grammars. In Dowty, D., Karttunen, L., and Zwicky, A. (editors), *Natural Language Processing—Theoretical, Computational and Psychological Perspectives.* Cambridge University Press, New York. (Originally presented in a Workshop on Natural Language Parsing at Ohio State University, Columbus, Ohio, May 1983)

Joshi, Aravind K. 1987. An Introduction to Tree Adjoining Grammars. In Manaster-Ramer, A. (editor), *Mathematics of Language.* John Benjamins, Amsterdam.

Kaplan, R., Netter, K., Wedekind, J., and Zaenen, A. 1989. Translation by structural correspondences. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL'89).* Manchester.

Kroch, Anthony and Joshi, Aravind K. 1985. *Linguistic Relevance of Tree Adjoining Grammars.* Technical Report MS-CIS-85-18, Department of Computer and Information Science, University of Pennsylvania.

Schabes, Yves, Abeillé, Anne, and Joshi, Aravind K. 1988. Parsing Strategies with 'Lexicalized' Grammars: Application to Tree Adjoining Grammars. In *Proceedings of the* 12^{th} *International Conference on Computational Linguistics (COLING'88).* Budapest, Hungary.

Shieber, Stuart and Schabes, Yves. 1990. Synchronous Tree Adjoining Grammars. in *Proceedings of the* 13^{th} *International Conference on Computational Linguistics (COLING'90).* Stockholm, Sweden.