Representing Paraphrases Using Synchronous Tree Adjoining Grammars

Mark Dras

Microsoft Research Institute
Macquarie University
Sydney. NSW. Australia. 2109.
markd@mpce.mq.edu.au

Summary

This paper looks at representing paraphrases, in the context of Reluctant Paraphrasing, using the formalism of Synchronous TAGs; it looks particularly at comparisons with machine translation and the modifications it is necessary to make to Synchronous TAGs for paraphrasing.

1. Introduction

This choosing of a representation for paraphrases is carried out in the context of what will be termed in the rest of this paper Reluctant Paraphrasing (RP). This is to contrast it with the sort of paraphrasing carried out by, for example, style checkers (e.g. Cherry, 1981; Kieras, 1990). In style checking, a typical aim may be to produce a more "active" text, and the means for this is a passive-to-active transformation. In this sort of situation, whenever a passive voice sentence is encountered, a decision to transform it is taken. This kind of paraphrase occurs often in remedial contexts like style checking; and it can be thought of as both "local", in that the decision to paraphrase depends only on the sentence and nothing else, and also as "eager", in that this same decision is taken at every opportunity. By contrast, Reluctant Paraphrasing is carried out on texts which are considered to accurately convey the writer's meaning, but which have to be adjusted for a set of overall external textual constraints, such as reduced text length or an improved score on a readability metric. Paraphrasing in this context is carried out "globally" - for example, only some passive voice sentences may be changed to active to shorten the text, depending on what is necessary to satisfy the given constraints.

"Local" paraphrasing implies the need for an arbitrarily complex representation formalism, as shortcomings to be corrected by remedial paraphrasing are arbitrarily complex. The shortcomings range from being syntactic in nature, such as "too much passive voice", to the much more abstract, such as "incoherent theme"; having such a wide range of abstractions makes the development of a formalism which will cover all of them a formidable task. However, under RP the paraphrases are only a means to an end - so, if constraints are extended to be more complex than just compressing a text, the set of paraphrases does not need to be likewise extended in terms of complexity - so the representation does not have to handle paraphrases of arbitrary complexity. The idea in this paper is thus to use a fairly simple representation, that of Synchronous TAGs (Shieber and Schabes, 1990), to describe paraphrasing; the appropriateness of Synchronous TAGs as a representation formalism is discussed, and modifications suggested based on differences between their previous applications - particularly concentrating on machine translation, given that it is also a text-to-text mapping process - and use in paraphrasing.

2. Paraphrases

The paraphrases used under RP, and discussed in this paper, are taken from commentators on stylistic paraphrase, both popular (Strunk and White, 1979; Kane, 1983) and academic (Jordan, 1993): this stylistic type of paraphrase was chosen because it keeps the meaning of

the text substantially the same, as well as because of their claims to "improve" text[1]. A sample set of such paraphrases is listed below.

(1) *Sarah, who put the book in the bed, was eating an apple.*
Sarah was eating an apple. Sarah put the book in the bed.

(2) *The apple was broken by the artist.*
The artist broke the apple.

(3) *Sarah ate the apple chosen by the artist.*
Sarah ate the apple. It had been chosen by the artist.

(4) *It was the artist who broke the apple.*
The artist broke the apple.

(5) *Sarah ate the apple which was chosen by the artist.*
Sarah ate the apple chosen by the artist.

(6) *Nigel smelled the familiar flowers.*
Nigel smelled the flowers.

The paraphrases described in this work are almost purely syntactic: that is, they can be represented in terms of a mapping between syntax trees describing each of the paraphrase alternatives.

3. Choice of Representation

A paraphrase representation can be thought of as comprising two parts - a representation for each of the source and target texts, and a representation for mapping between them. Tree Adjoining Grammars (TAGs) cover the first part: as a formalism for describing the syntactic aspects of text, they have a number of desirable features. The properties of the formalism are well established, and have been explored extensively: TAGs have been shown to be useful for natural language work, as they are capable of describing non-context-free aspects of natural languages, without being too powerful (Joshi et al, 1975). The research has also led to the development of a large standard grammar (XTAG Research Group, 1995), minimising the amount of grammar development work involved; this also reduces the risk that a grammar will be developed which has been moulded to facilitate paraphrasing. Using an existing grammar reduces the chance for short-cuts based on a sympathetic representation. Related to this, XTAG, a parser complete with GUI, has been built for TAGs (Doran et al, 1994). This is especially useful when building an actual system, as the parser can be plugged in as a component.

Mapping between source and target texts is achieved by an extension to the TAG formalism known as Synchronous TAG, introduced by Shieber and Schabes (1990). Synchronous TAGs (STAGs) comprise a pair of trees, at least one of which is a TAG tree, plus links between nodes of the trees. The original paper of Shieber and Schabes proposed using synchronous TAGs to map from a syntactic to a semantic representation, while another paper by Abeillé (1990) proposed their use in machine translation. The first of these uses allows the extension of Reluctant Paraphrasing to a broader domain, that of semantic paraphrasing: currently in RP, as mentioned, only syntactic paraphrases (that is, a direct mapping involving one synchronous TAG pair) are handled; but having the same mechanism for translating to a semantic representation allows for semantic paraphrases

[1] The issue of to what extent meaning is preserved, and whether the difference in meaning can be quantified for optimisation purposes when choosing paraphrases, is the subject of future work.

also (that is, an indirect mapping involving two synchronous TAGs). The use in machine translation is quite close to the use proposed here; instead of mapping between possibly different trees in different languages, there is a mapping between trees in the same language with very different syntactic properties.

Note that this is by no means the only possible representation for paraphrasing: Transformational Generative Grammar (TGG), for example, was built around a notion of paraphrase; however, STAGs combine standardisation, parser availability, well-definedness and (unlike TGG) relative lack of controversy to make a good representation formalism.

4. Mappings

As mentioned in the previous section, the use of STAGs in paraphrasing is quite close to their use in machine translation. This section will look at the arguments Abeillé (1990) makes for the suitability of STAGs in translation, and investigate their applicability to paraphrasing; modifications to the formalism will be suggested based on this.

Abeillé notes that the STAG formalism allows an explicit semantic representation to be avoided, mapping from syntax to syntax directly. This fits well with the syntactic paraphrases currently used in RP; but it does not, as Abeillé also notes, preclude semantic-based mappings, with Shieber and Schabes constructing syntax-to-semantics mappings as the first demonstration of STAGs. Similarly, more semantically-based paraphrases (for example, *Sarah swallowed the mango chunks* to *Sarah ate the tropical fruit*) are possible through an indirect application of STAGs to a semantic representation, and then back to the syntax. In addition to Abeillé's reasoning for avoiding a semantic representation - that Occam's razor suggests the simplest form as a first attempt at explanation, and makes it necessary to justify any more complex additions - this syntactic representation has been chosen for RP for a number of reasons. Most important is that paraphrases suggested by commentators on stylistics (e.g. Kane, 1983; Jordan, 1993) and technical writing experts (Klare, 1979) are often syntactic in nature; while there are other, more general, recommendations, such as "Be concise" from Strunk and White (1979), the syntactic paraphrases are the simplest to apply and the least ambiguous.

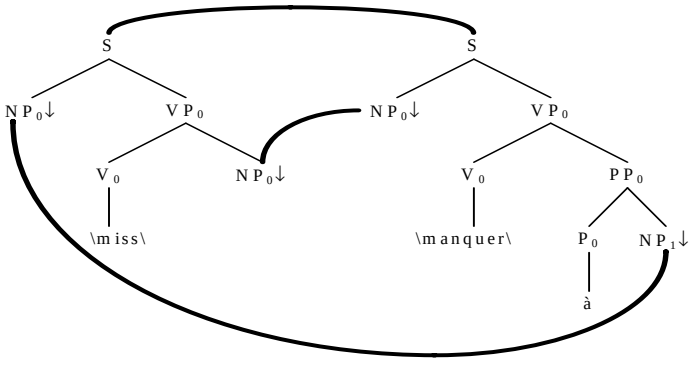

Figure 1: STAG representation for English-French *miss-manquer à*.

One major difference is in lexicalisation. The sorts of mappings that Abeillé deals with are lexically idiosyncratic: the English sentences *Kim likes Dale* and *Kim misses Dale*, while syntactically parallel and semantically fairly close, are translated in different ways into French - to a similar simple transitive verb structure in the former case (*Kim aime Dale*) but to a different structure in the latter (*Dale manque à Kim*). The actual STAG for the *misses*

case is given in Figure 1[2]. The actual mappings depend on the properties of words, so any TAGs used in this synchronous manner will necessarily be lexicalised. In RP, however, the sorts of paraphrases which are used are lexically general: splitting off a long noun post-modifier, as in (3), is not dependent on the noun, its modifier, or any other lexical element in the sentence.

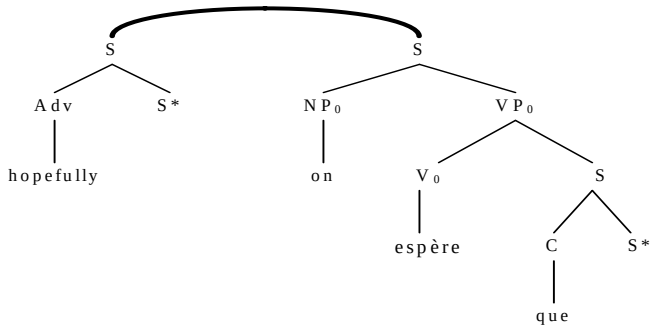

Figure 2: STAG representation for English-French hopefully-on espère que.

Related to this is that, at least between English and French, extensive syntactic mismatch is unusual, much of the difficulty in translation coming from lexical idiosyncrasies. A consequence, for machine translation, is that much of the synchronising of TAGs is between elementary trees. So, even with a more complex syntactic structure than the translation examples above - for example, *Kim, who is now the saddest boy in the class, missed Dale* to *Dale manque à Kim, qui est le garçon le plus triste dans la classe* - the changes can be described in terms of mappings between elementary trees, or just in the transfer lexicon: the *miss*\ to *manquer à*\ tree, mentioned previously; the tree covering post-modification of nouns by adjectives, reversed for pre-modification; superlative forms of adjective X becoming *le/la plus* X. Abeillé notes that there are occasions where it is necessary to replace an elementary tree by a derived tree; for example, when the number of arguments in the predicate being translated does not match that of the target language predicate. The example she gives is *Hopefully, John will work* becoming *On espère que Jean travaillera*; the translation of *hopefully* can be represented by a TAG tree pair, where the elementary auxiliary tree for the English *hopefully* is paired with the derived tree for the French *on espère que*. This is displayed in Figure 2.

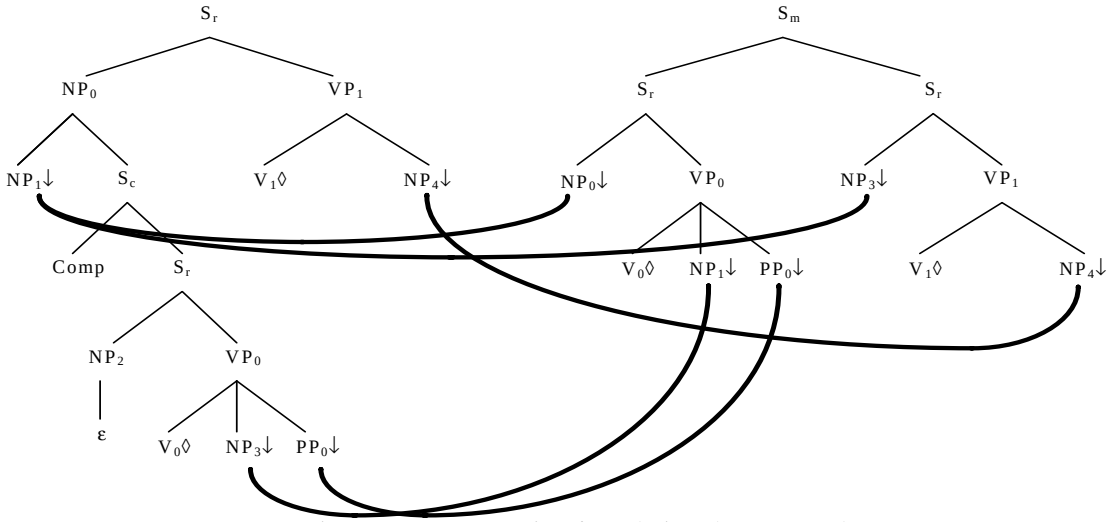

Figure 3: representation for relative clause paraphrase.

[2] Standard TAG notation, in line with Abeillé (1990), is used: '↓' represents nodes to be substituted, '*' annotates the foot node of an auxiliary tree, *word*\ denotes the morphological variants of *word*. The trees are combined with substitution and adjunction.

This occurs much more frequently in RP: by definition, the mappings are between units of text with differing syntactic properties. For example, the mapping of (1) involves the pairing of two derived trees, as in Figure 3.

In this case, both trees are derived ones. A problem with the STAG formalism in this situation is that it doesn't capture the generality of the mapping between (1a) and (1b); separate tree pairings will have to be made for verbs in the matrix clause which have complementation patterns different from that of the above examples: intransitive verbs, object-equi verbs, and so on; the same is true for verbs in the subordinate clause. For more complex matchings, the making and pairing of derived trees becomes combinatorially large.

A more compact definition is to have links, of a kind different from the standard STAG links, between nodes higher in the tree. In STAG, a link between two nodes specifies that any substitution or adjunction occurring at one node must be replicated at the other. This new proposed link would be a summary, synchronise-this-entire-subtree link: more precisely, each subnode of the node with the summary link is mapped to the corresponding node in the paired tree in a synchronous depth-first traversal of the subtree. Naturally, this can only be defined for pairs of nodes which have the same structure[3]; that is, in the context of paraphrasing, it is effectively a statement that the paired subtrees are identical. So, for example, a mapping between the nodes labelled VP_1 in each of the trees of the example described above would be an appropriate place to have such a summary link: by establishing a mapping between each subnode of VP_1, this covers the different types of matrix clauses mentioned above: intransitive, ditransitive, etc. In Figure 4, these summary links are represented by the thick dashed line.

Another feature of using STAGs for paraphrasing is that the links are no longer necessarily one-to-one. In the right-hand tree of the Figure 3 pairing, the subject NPs of both sentences are linked to NP_1 of the left-hand tree; this is a statement that both resulting sentences have the same subject. This does not, however, change the properties in any significant way.

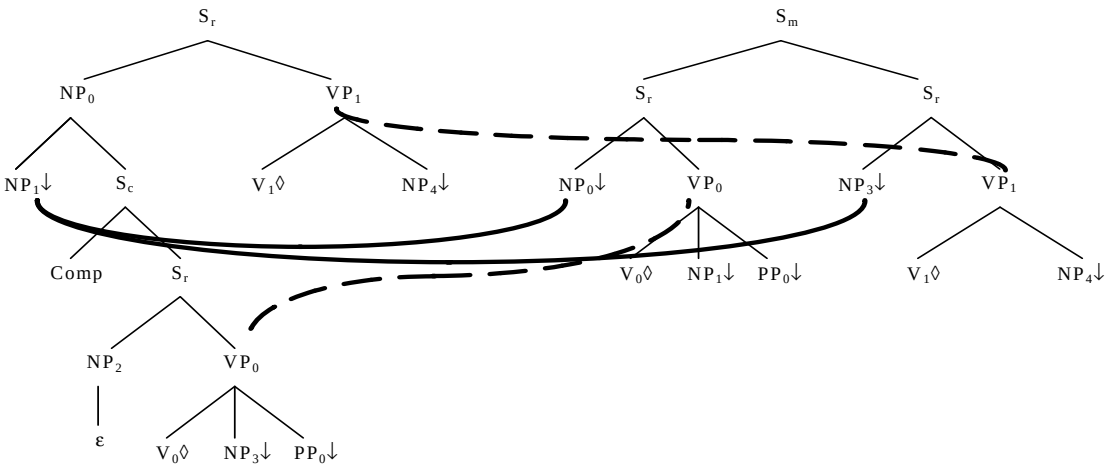

Figure 4: representation with summary links for relative clause paraphrase.

It is also useful to add another type of link which is non-standard, in that it is not just a link between nodes at which adjunction and substitution occur, but which represents shared attributes. It connects nodes such as the main verb of each tree, and indicates that particular attributes are held in common. For example, mapping between active and

[3] More precisely, they need only have the same number and type of argument slots.

passive voice versions of a sentence, as in (2), is represented by the tree in Figure 5. The verb in the active version (*breaks*) shares the attribute of tense with the auxiliary verb *be*\ (*was*), and the lexical component is shared with the main verb of the passive tree (*broken*), which takes the past participle form. This sort of link is unnecessary when STAGs are used in MT, as the trees are lexicalised, and, according to Abeillé, the information is shared in the transfer lexicon. Her example is the translation of *will work* to *travaillera*: French has no word corresponding to the modal *will*, which is represented rather as a flexion, attached here to the infinitive verb *travailler*; and this correspondence is described in the transfer lexicon. Since, under RP, the transfer lexicon does not play such a role, the shared information is represented by this new type of link between the trees, where the links are labelled according to the information shared. Hence, node V_1 in the active tree has a TENSE link with node V_0 in the passive tree, where tense is the attribute in common; and a LEX link with node V_1 in the passive tree, where the lexeme is shared. The determination of a precise set of link labels is future work.

5. Notation

In RP, the tree notation thus becomes fairly clumsy: as well as consuming a large amount of space (given the large derived trees), it fails to reflect the generality provided by the summary links. That is, it is not possible to define a mapping between two structures reflecting their common features if the structures are not, as is standard in STAG, entire elementary or derived trees. Therefore, a new and more compact notation is proposed to overcome these two disadvantages.

The new notation consists of three parts: the first part uniquely defines each tree of a synchronous tree pair; the second part describes, also uniquely, the nodes that will be part of the links; the third part links the trees via these nodes. So, let variables *X* and *Y* stand for any string of argument types acceptable in tree names; so, for example, *X* could be nx1nx2 and *Y* n1, or *X* could be nx1pnx2 and *Y* nx1nx2. Then, for example, the tree for (1a) can be defined as the adjunction of a βN0nx0V*X* tree (generic relative clause tree, standing for βN0nx0V nx1nx2, or βN0nx0V nx1pnx2, etc) into an αn0V*Y* tree; the tree for (1b) can be defined as a conjoined S tree, having a parent S_m node and 2 child nodes αn0V*X* and αn0V*Y*.

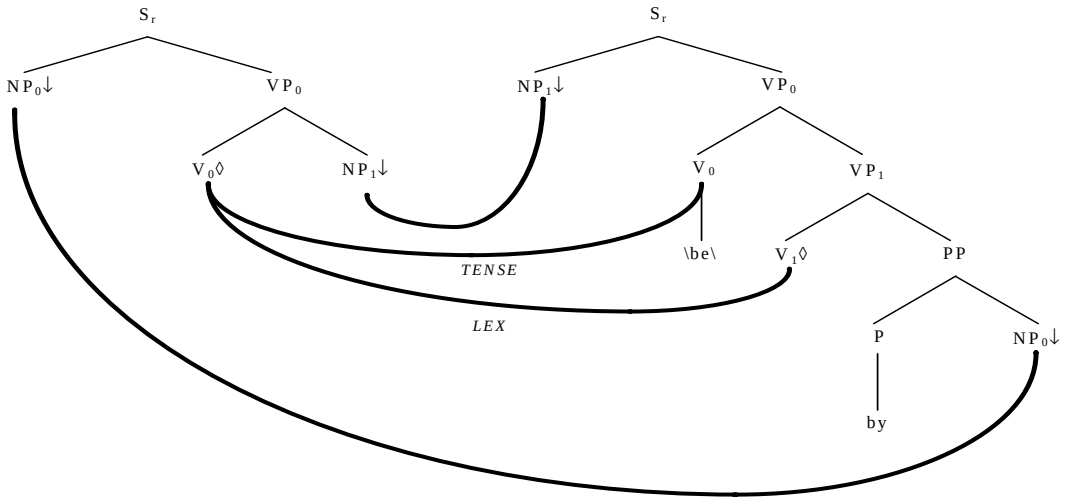

Figure 5: representation with non-standard links for active-passive paraphrase.

The second part of the notation requires picking out important nodes. For describing these nodes, the numbering scheme used in the diagrams is inadequate - numbering generally

follows a depth-first pattern, meaning that, in Figure 4, the numbering of NP nodes within the matrix clause VP is dependent on the number of NP nodes in the adjoined βN0nx0Vnx1pnx2 tree. Thus, a different identification scheme needs to be used, such as describing nodes in terms of position with respect to other nodes, in such a manner that the key node is identified uniquely[4]. One possibility is XTAG's internal Lisp representation for nodes, which gives the label of a node, generally its grammatical category, along with parent and children nodes. In this representation, the node NP1 from Figure 4 could be identified as

(label:NP parent:(label:NP parent:(label:S_r parent:NIL)))

A more concise version of this could be a string comprising node labels with relations between them, signifying a relationship taken from the set {parent, child, left-sibling, right-sibling}, with the abbreviation {p, c, ls, rs}. The node can then be described by the string NPpNPpS_rpNIL. In this manner, the important nodes of Figure 4 can be described by strings and associated mnemonic nicknames:

	NODE	STRING	NICKNAME
lefthand tree	NP_1	NPpNPpS_rpNIL	T_1subjNP
	VP_1	VPpS_rpNIL	T_1matrixVP
	VP_0	VPp$S_r$lsComp	T_1embeddedVP
righthand tree	NP_0	NPp$S_r$rsS_r	T_2leftsubjNP
	VP_0	VPp$S_r$rsS_r	T_2leftVP
	NP_3	NPp$S_r$lsS_r	T_2rightsubjNP
	VP_1	VPp$S_r$rsS_r	T_2rightVP

The third part of the representation is then linking the nodes. Standard links are represented by an equal sign; other links are represented with the link type subscripted to the equal sign. Thus, for Figure 4:

T_1subjNP = T_2leftsubjNP
T_1subjNP = T_2rightsubjNP
T_1matrixVP $=_{SUMMARY}$ T_2leftVP
T_1embeddedVP $=_{SUMMARY}$ T_2rightVP

For the active-passive mapping of Figure 5, the representation would thus be

T_1: tree αnx0Vnx1
nodes (nickname and i.d. string)
T_1subjNP NPpS_r
T_1mainV V◊
T_1objNP NPpVP

T_2: tree αnx1Vbynx0 with adjoined βVvx *be*\
nodes (nickname and i.d. string)
T_2subjNP NPpS_r
T_2auxV VrsVPcV◊
T_2mainV V◊
T_2objNP NPlsPc"by"

[4] This uniqueness includes the ensuring that the positions are still uniquely defined if a substitution or adjunction occurs; that is, if the nodes are defined relative to each other, the relative placement does not change, nor does the definition lose its unique status. Showing that this is the case will be future work.

Equations:

T_1subjNP	=	T_2objNP
T_1mainV	$=_{TENSE}$	T_2auxV
T_1mainV	$=_{LEX}$	T_2mainV
T_1objNP	=	T_2subjNP

6. Conclusion

Synchronous TAGs are a useful representation for paraphrasing - mapping between texts of the same language but with different syntactic structure - in the context of Reluctant Paraphrasing. A number of modifications need to be made, however, to properly capture the nature of paraphrases: the creation of a new type of summary link, to compensate for the increased importance of derived trees; the allowing of many-to-many links between trees; the creation of partial links, which allow some information to be shared; and a new notation which expresses the generality of paraphrasing.

7. References

Abeillé, Anne, Y. Schabes and A. Joshi (1990). "Using Lexicalised Tags for Machine Translation". In *Proceedings of COLING90*, 1-6.

Cherry, Lorinda (1981). *Writing Tools - The STYLE and DICTION Programs*. Bell Labs Computing Science Technical Report No. 91.

Doran, Christy, D. Egedi, B.A. Hockey, B. Srinivas and M. Zaidel (1994). "XTAG System - A Wide Coverage Grammar of English". In *Proceedings of COLING94*, 922-928.

Jordan, Michael (1993). "Towards an Understanding of Mature Writing: Analyzing and Paraphrasing Complex Noun Phrases". *Technostyle*, 11(2), 39-72.

Kane, Thomas S (1983). *The Oxford Guide to Writing*. Oxford University Press. New York, NY.

Kieras, David (1990). *The Computerized Comprehensibility System Maintainer's Guide*. University of Michigan Technical Report no. 33.

Klare, George (1979). *Readability Standards for Army-wide Publications (Evaluation Report 79-1)*. US Army Administrative Centre. Fort Benjamin Harrison, IN.

Joshi, Aravind, L. Levy and M. Takahashi (1975). "Tree Adjunct Grammars". *Journal of Computer and System Sciences*, 10(1).

Strunk, William and E. B. White (1979). *The Elements of Style, 3rd edition*. MacMillan Publishing Co. New York, NY.

Shieber, Stuart and Y. Schabes (1990). "Synchronous Tree Adjoining Grammars". In *Proceedings of COLING90*, 253-258.

XTAG Research Group (1995). *A Lexicalised Tree Adjoining Grammar for English*. University of Pennsylvania Technical Report IRCS 95-03.