Synchronous Structured String-Tree Correspondence (S-SSTC)

MOSLEH H. AL-ADHAILEH
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, MALAYSIA
mosleh@cs.usm.my, mosleh@hotmail.com

TANG ENYA KONG
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, MALAYSIA
enyakong@cs.usm.my

ABSTRACT

In this paper, a flexible annotation schema called Structured String-Tree Correspondence (SSTC) is introduced. We propose a variant of SSTC called synchronous SSTC. Synchronous SSTC can be used to describe the correspondence between different languages. We will also describe how synchronous SSTC provides the flexibility to treat some of the non-standard cases, which are problematic to other synchronous formalisms. The proposed synchronous SSTC schema well suited to describe the correspondence between different languages, in particular, relating a language with its translation in another language (i.e. in Machine Translation). Synchronous SSTC can be used as annotation for translation systems that automatically extract transfer mappings (rules or examples) from bilingual corpora. The synchronous SSTC also can be used to construct a Bilingual Knowledge Bank (BKB), where the examples are kept in form of synchronous SSTCs.

KEYWORDS: Natural Language Processing (NLP), parallel text, Structured String-Tree Correspondence (SSTC), Synchronous SSTC, Bilingual Knowledge Bank (BKB).

1. INTRODUCTION

In this paper, a flexible annotation schema called Structured String-Tree Correspondence (SSTC) [3] will be introduced to capture a natural language text, its corresponding abstract linguistic representation and the mapping (correspondence) between these two. The correspondence between the string and its associated representation tree structure is defined in terms of the sub-correspondence between parts of the string (substrings) and parts of the tree structure (subtrees), which can be interpreted for both analysis and generation. Such correspondence is defined in a way that is able to handle the non-standard cases (non-projective correspondence).

In order to describe the relation between different languages, we will define a variant of SSTC called synchronous SSTC. Synchronous SSTC consists of two SSTCs that are related by synchronization relation. The use of synchronous SSTC is motivated by the desire to describe not only the correspondence between the text and its representation structure for each language (i.e. SSTC) but also the correspondence between two languages (synchronous correspondence). For instance, between a language and its translation in other language, the case of Machine Translation. The synchronous SSTC will be used to relate expression of a natural language to its associated translation in another language. The interface between the two languages is made precise via the synchronization relation between two SSTCs, which is totally non-directional.

In this paper, we will present the proposed synchronous SSTC – a schema well suited to describe the correspondence between two languages. The synchronous SSTC is flexible and able to handle the non-standard correspondence cases exist between different languages. It can also be used to facilitate automatic extraction of transfer mappings (rules or examples) from bilingual corpora.

2. STRUCTURED STRING-TREE CORRESPONDENCE (SSTC)

From the Meaning-Text Theory (MTT)[1] point of view, Natural Language (NL) is considered as a correspondence between meanings and texts [6]. The MTT point of view, even if it has been introduced in different formulations, is more or less accepted by the whole linguistic community.

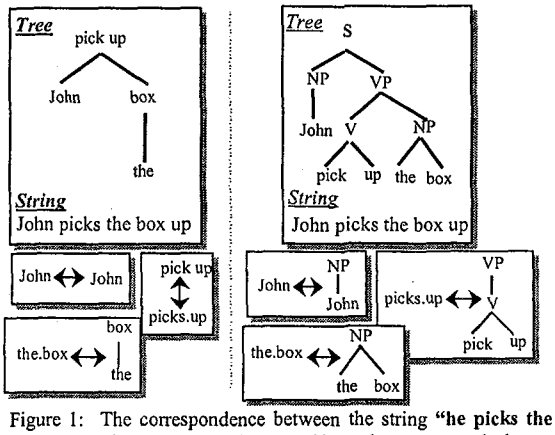

Figure 1: The correspondence between the string **"he picks the box up"** and its representation tree (dependency tree and phrase-structure tree), together with the sub-correspondence between the substrings and subtrees.

[1] The Meaning-Text Theory (MTT) was put forward in [15], in the framework of research in Machine translation. More presentations of MTT can be found in [7] and [8].

However, NL is not only a correspondence between different representation levels, as stressed by MTT postulates, but also a sub-correspondence between them. For instance, between the string in a language and its representation tree structure, it is important to specify the correspondence between the string and its associated tree structure, and more than that the sub-correspondence between parts of the string (substrings) and parts of the tree structure (subtrees), which can be interpreted for both analysis and generation in NLP. Also it is important to define the correspondence in a way that is able to handle the non-standard cases (non-projective correspondence), see the example in Figure 1. It is well known that many linguistic constructions are not projective (e.g. scrambling, cross serial dependencies, etc.). In this section, we attempt to introduce a flexible annotation structure called Structured String-Tree Correspondence (SSTC). We stress on the fact that in order to describe Natural Language (NL) in a natural manner, three distinct components need to be expressed by the annotation structure; namely, the text, its corresponding abstract linguistic representation and the mapping (correspondence) between these two.

2.1 The SSTC Annotation Structure

The SSTC is a general structure that can associate an arbitrary tree structure to string in a language as desired by the annotator to be the interpretation structure of the string, and more importantly is the facility to specify the correspondence between the string and the associated tree which can be non-projective [3]. These features are very much desired in the design of an annotation scheme, in particular for the treatment of linguistic phenomena, which are non-standard, e.g. crossed dependencies [14].

In the SSTC, the correspondence between the sentence on one hand, and its representation tree on the other hand, is defined in terms of finer sub-correspondences between substrings of the sentence and subtrees of the tree. Such correspondence is made of two interrelated correspondences, one between nodes and substrings, and the other between subtrees and substrings, (the substrings being possibly discontinuous in both cases). The notation used in SSTC to denote a correspondence consists of a pair of intervals X/Y attached to each node in the tree, where X(SNODE) denotes the interval containing the substring that corresponds to the node, and Y(STREE) denotes the interval containing the substring that corresponds to the subtree having the node as root [3].

Definitions*[2]*:

- *An **SSTC** is a general structure, which is a string in a language associated with an arbitrary tree structure; i.e. its interpretation structure, and the correspondence between the string and its associated tree, which can be non-projective; i.e. SSTC is a triple (**st**, **tr**, **co**), where **st** is a string in one language, **tr** is its associated representation tree structure and **co** is the correspondence between **st** and **tr**.*
- *The correspondence **co** between a string and its representation tree is made of two interrelated correspondences:*
  - *a) Between nodes and substrings (possibly discontinuous).*
  - *b) Between (possibly incomplete) subtrees and (possibly discontinuous) substrings.*
- *The correspondence can be encoded on the tree by attaching to each node N in the representation tree two sequences of **INTERVALS** called **SNODE(N)** and **STREE(N)**.*
- ***SNODE(N)**: An interval of the substring in the string that corresponds to the node N in the tree.*

  ***STREE(N)**: An interval of the substring in the string that corresponds to the subtree having the node N as a root in the tree.*

  ***SNODE** and **STREE** intervals are attached to each node in the representation tree.*

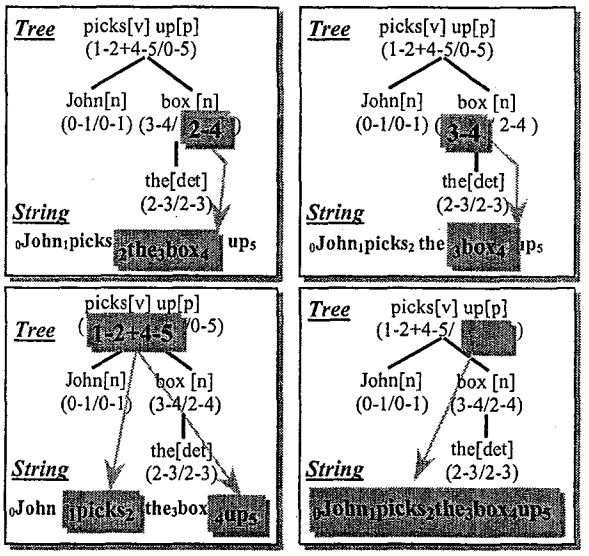

Figure 2: An SSTC recording the sentence "**John picks the box up**" and its dependency tree together with the example correspondences between substrings of the sentence and subtrees of the tree.

Figure 2 illustrates the sentence **"John picks the box up"** with its corresponding SSTC. It contains a non-projective correspondence. In the dependency structure tree, an interval is assigned to each word in the sentence, i.e. (0-1) for **"John"**, (1-2) for **"picks"**, (2-3) for **"the"**, (3-4) for **"box"** and (4-5) for **"up"**. A substring in the sentence that corresponds to a node in the representation tree is denoted by assigning the interval of the substring to SNODE of the node, e.g. the node **"picks up"** with SNODE intervals (1-2+4-5) corresponds to the words **"picks"** and **"up"** in the string with the similar intervals, the node **"box"** with SNODE interval (3-4) corresponds to the word **"box"** in the string with the similar interval. The correspondence between subtrees and substrings are denoted by the interval assigned to the STREE of each node[3], e.g. the subtree rooted at node **"picks up"** with STREE interval (0-5) corresponds to the whole sentence **"John picks the**

[2] These definitions are based on the discussion in [12] and [3].

[3] For the computation of the String-Tree correspondences, see [12].

box up", the subtree rooted at node "**box**" with STREE interval (2-4) corresponds to the phrase "**the box**" in the string. In the phrase structure tree, the same notation (as in the dependency structure tree) is used to denote the correspondence. Note that in phrase structure tree all internal nodes that do not correspond to any word in the sentence are denoted by the assignment of ϕ (interval of length 0) to the (respective) SNODE (see Figure 3).

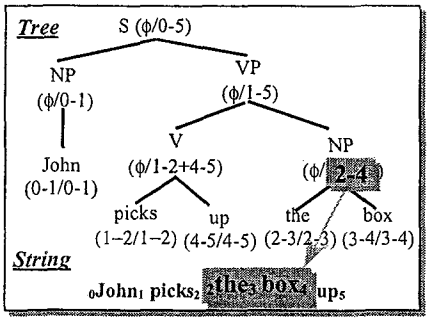

Figure 3: An SSTC recording the sentence "**John picks the box up**" and its phrase structure tree together with the correspondences between substrings of the sentence and subtrees of the tree.

One can easily imagine the case where in the representation tree (TREE), elements of the string may have been erased, duplicated, or transposed to different positions with respect to the order in the STRING, while some discontinuous groups may have been put together to form adjacent constituents. As we mentioned earlier, a flexible annotation structure should be able to handle these non-standard cases. The case depicted in Figure 2, describes how the SSTC structure treats some non-standard linguistic phenomena. The particle **"up"** is featurised into the verb **"pick"** and in discontinuous manner (e.g. **"up" (4-5)** in **"pick-up" (1-2+4-5)**) in the sentence **"He picks the box up"**.

In general, SSTC can be used for specifying the correspondence between a text and its associated representation structure in such a way that the correspondence can be interpreted for writing linguistic programs for both analysis and synthesis in most Machine Translation systems. The same properties also mean that SSTC is independent of any particular linguistic theory, and so any theory adopts an equivalent type of data structure and does not make use of procedural mechanisms to explain linguistic phenomena should be able to use the SSTC as a notational support for the theory. For more on proprieties of SSTC see [3]

3. SYNCHRONOUS STRUCTURED STRING -TREE CORRESPONDENCE (S-SSTC)

Much of theoretical linguistic can be formulated in a very natural manner as stating correspondences (translations) between layers of representation structures [9]. An analogous problem is to be defined in such a way that expresses structural correspondences between representation trees, for instance, the correspondence between a language and its translations in other languages. Therefore the synchronization of two structures or two adequate linguistic formalisms seems to be an appropriate representation for that.

The idea of parallelized formalisms is widely used one, and one which has been applied in many different ways. The use of synchronous formalisms is motivated by the desire to describe two languages that are closely related to each other but that do not have the same structures. This is for example the case in machine translation or for the relation between syntax and semantics.

Synchronous Tree Adjoining Grammar (S-TAG) is a variant of Tree Adjoining Grammar (TAG) introduced by [11] to characterize correspondences between tree adjoining languages. They can be used to relate TAGs for two different languages, for example, for the purpose of immediate structural translation in machine translation [1], [4], or for relating a syntactic TAG and semantic one for the same language [11].

Considering the original definition of S-TAGs, one can see that it does not restrict the structures that can be produced in the source and target languages, i.e. it allows the construction of a non-Tree-Adjoining-Language [10], [5]. As a result, [10] proposed a restricted definition for S-TAG, namely, the IS-TAG for isomorphic S-TAG. In this case only Tree-Adjoining-Languages can be formed in each component. This isomorphism requirement is formally attractive, but for practical applications somewhat too strict. Also contrastive well-known translation phenomena exist in different languages, which cannot be expressed by isomorphic S-TAG, Figure 4 illustrates some examples.

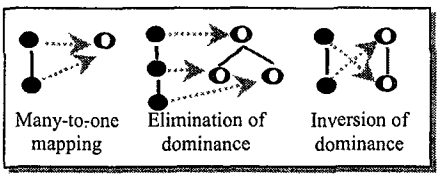

Figure 4: Kinds of relations between different languages.

In general, it is difficult to decide whether synchronization of two TAGs or rather any two grammars is the right approach for expressing structural correspondences. More importantly is to have links, of kind different from the standard STAG links, between nodes higher in the tree. Links that explicitly define the correspondences at different levels of the synchronized structures. Since the analysis and the generation task require more than context-free formalisms, and more flexibility in defining the correspondences, we propose a flexible annotation schema (i.e. Synchronous Structured String-Tree Correspondence (S-SSTC)) to realize additional power and flexibility in expressing structural correspondences.

3.1 The Synchronous SSTC

Synchronous SSTC consists of two SSTCs that are related by synchronization relation. As discussed previously, the

attractive way SSTC describing NL and its flexibility to manifest the correspondence between string in a natural language and its representation tree, present a challenge for NL applications, for instance, to the task of automatic translation of natural language (MT). Basic linguistic structures must be created in such a way that expresses structural correspondences at different levels of the structures; therefore the synchronization of two SSTCs seems to be an appropriate representation for that.

Definitions:

- *Let S and T be a triple SSTCs (st, tr, co), where st is a string in one language, tr is its associated representation tree structure and co is the correspondence between st and tr, as defined in Section 2.3.*
- *A synchronous SSTC S_{syn} is a triple $(S, T, \varphi_{(S,T)})$, where $\varphi_{(S,T)}$ is a set of links, which defines the synchronous correspondences between the nodes of tr in S, and the nodes of tr in T, at different internal levels of the two SSTC structures.*
- *For each elementary unit (i.e. node, subtree or partial subtree) N_s in the first SSTC S, there is N_t, $\{N_t\}$ or ε elementary unit/s in the second SSTC T correspond/s to it.*
- *Each pair (N_s, N_t) -where N_s corresponds to N_t has a synchronous correspondence link $l \in \varphi_{(S,T)}$ between them of type l_{sn} or l_{st}.*
- *A correspondence of l_{sn} type, means there is a synchronous correspondence between the node N_s and the node N_t of the specified unit (N_s, N_t).*
- *A correspondence of l_{st} type, means there is a synchronous correspondence between the subtree rooted by N_s and the subtree rooted by N_t of the specified unit (N_s, N_t).*
- *l_{sn} and l_{sn} synchronous correspondences can be between nodes and subtrees with non-standard phenomena; i.e. featursiation and discontinuity (crossed dependency), see Figures 5 and 9.*

The synchronous SSTC will be used to relate expressions of a natural language to its associated translation in another language. For convenience, we will call the two languages *source* and *target* languages, although synchronous SSTC is totally non-directional. Synchronous SSTC is defined to make such relation explicit. The *source-target* interface is made precise by using a synchronous SSTC, i.e. two SSTCs, one represents the *source* and the other represents the *target*, which are interfaced via synchronization relations. Figure 5 depicts a synchronous SSTC for the English *source* sentence ***"John picks the heavy box up"*** and its translation in the Malay *target* sentence ***"John kutip kotak berat itu"***. The gray arrows indicate the correspondence between the string and it representation tree within each of the SSTCs, and the dot-gray arrows indicate the relations (i.e. synchronous correspondence) of synchronization between linguistic units of the *source* SSTC and the *target* SSTC.

Based on the notation used in synchronous SSTC, Figure 5 illustrates the synchronous SSTC for the English sentence ***"John picks the heavy box up"*** and its translation in the Malay language ***"John kutip kotak berat itu"***, with the synchronous correspondence between them. The synchronous correspondence is denoted in terms of SNODE pairs for l_{sn} and STREE pairs for l_{st}. For l_{sn} each pair is of (SN_s, SN_t), where SN_s is SNODE interval in the *source* SSTC and SN_t is SNODE interval in the *target* SSTC. Also for l_{st} each pair is of (ST_s, ST_t), where ST_s is STREE interval in the *source* SSTC and ST_t is STREE interval in the *target* SSTC. For instance, as depicted in Figure 5, the fact that **"picks up"** in the *source* corresponds to **"kutip"** in the *target* is expressed by the pair $(SN_s, SN_t) \Leftrightarrow$(1-2+5-6,1-2) under the l_{sn} synchronous correspondence. Whereas, the fact that ***"John picks the heavy box up"*** is corresponds to ***"John kutip kotak berat itu"*** is expressed by $(ST_s, ST_t) \Leftrightarrow$(0-6,0-5) under the l_{st} synchronous correspondence. Also the fact that **"box"** in the *source* corresponds to **"kotak"** in the *target* under the pair $(SN_s, SN_t) \Leftrightarrow$(4-5,2-3) in the l_{sn} synchronous correspondence. Whereas, the phrase **"the heavy box"** is corresponds to the phrase **"kotak berat itu"** in the *target* is expressed by $(ST_s, ST_t) \Leftrightarrow$(2-5,2-5) under the l_{st} synchronous correspondence.

1) ***English:*** John picks the heavy box up
   Malay: John kutip kotak berat itu

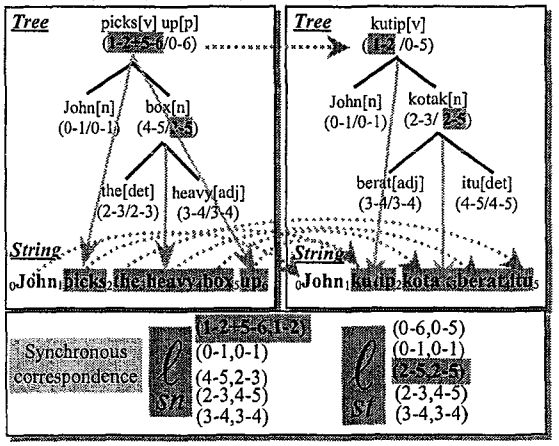

Figure 5: A synchronous SSTC for the sentence **"John picks the heavy box up"** and its Malay translation **"John kutip kotak berat itu"**, with the synchronous correspondence between them.

4. HANDLING NON-STANDARD CASES WITH S-SSTC

As mentioned earlier, there are some non-standard phenomena exist between different languages, that cause challenges for synchronized formalisms. In this Section, we will describe some example cases, which are drawn from the problem of using synchronous formalisms to define translations between languages; i.e. modeling the synchronous correspondences at different levels between them. Some of these examples are taken from [10].

In the sentence pair (1) (see Figure 5), the *English* sentence has non-standard cases of featurisation and crossed dependency in **"picks up"**. A many-to-one synchronous correspondence where the words **"picks"** and **"up"** in the *source* correspond to **"kutip"** in the *target*. Another case is reordering of words in the phrases, which

is clear in the phrase "**the**$_{det}$ **heavy**$_{adj}$ **box**$_{n}$" and it corresponding phrase "**kotak**$_{n}$ **berat**$_{adj}$ **itu**$_{det}$" in the *target*.

Sentence pair (2) (see Figure 6) shows two non-standard cases between languages; e.g. *French* and *English*. First, the case of many-to-one correspondence, where a word (single node) in one language corresponds to a phrase (subtree) in the other, namely, the adverbial "**hopefully**" is translated by the *French* phrase "**On espére que**". Second, a case of argument swap (reordering of subtrees) in the *English* "**Kim misses Dale**" and its corresponding translation "**Dale manqué a Kim**" in *French*.

An even more extreme relationship between the synchronized pair involving inverted domination correspondences, is exemplified in sentence pair (3), Figure 7. In this case, the phrase "**en courant**" is an adverbial modifier to the verb "**monte**". Presumably, "**en courant**" would be corresponds to the English "**runs**" and "**monte**" with the English "**up**", at least under the most natural analysis. In the corresponding English SSTC the domination is inverted, where "**up**" is dominated be "**runs**".

The case described in sentence pair (4) (see Figure 8) exemplifies a case where the number of nodes in the synchronized SSTCs or subSSTCs is the same, but they exhibit different structures. Nodes participating in the domination relationship in one SSTC may be mapped to nodes neither of which dominates the other. The clitic "**lui**" in the French SSTC dominated by "**soigné**", although its corresponding pronoun "**his**" in English part is dominated by the object "**teeth**".

The sentence in (5), Figure 9, describes the cases of clitic climbing in French and the non-projective correspondence. It shows the flexibility of SSTC and synchronous SSTC in handling such popular cases.

2) ***French:*** On espére que Dale manqué a Kim.
English: Hopefully Kim misses Dale.

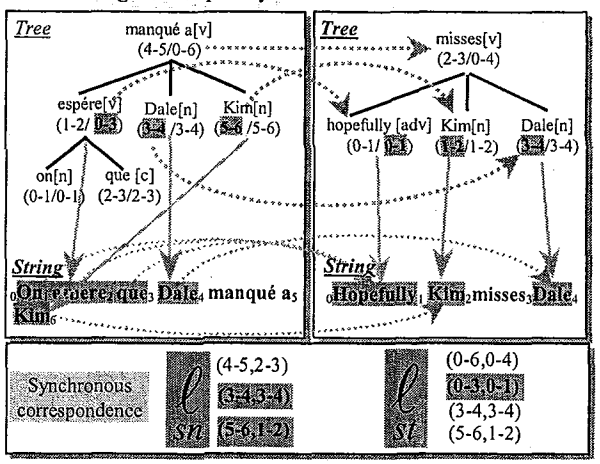

Figure 6: Many-to-one correspondence and arguments swapping correspondence in the *French* sentence "**On espére que Dale manqué a Kim**" and its corresponding *English* sentence "**Hopefully Kim misses Dale**".

3) ***French:*** Jean monte la rue en courant.
English: John runs up the street.

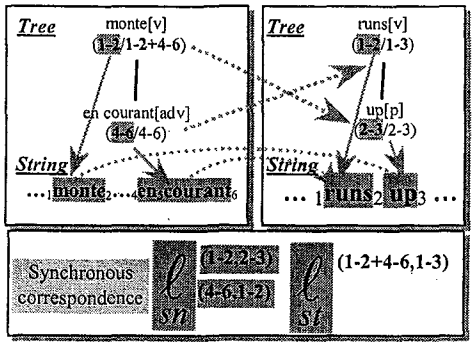

Figure 7: Inversion of dominance in the *French* sentence "**Jean monte la rue en courant**" and its corresponding *English* sentence "**John runs up the street**".

4) ***French:*** Le docteur lui soigné les dents.
English: The doctor treats his teeth.

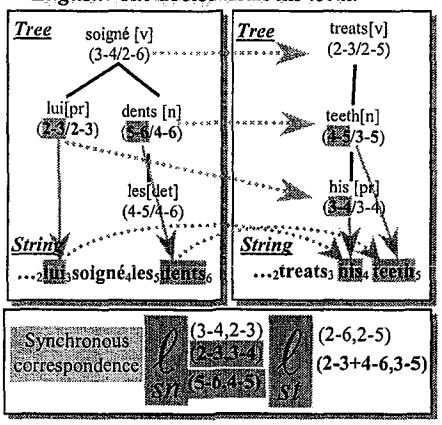

Figure 8: Elimination of dominance, in the French sentence "**le docteur lui soigné les dents**" and its corresponding English sentence "**the doctor treats his teeth**".

5) ***French:*** Pierre ne l 'a pas vu.
English: Peter has not seen it.

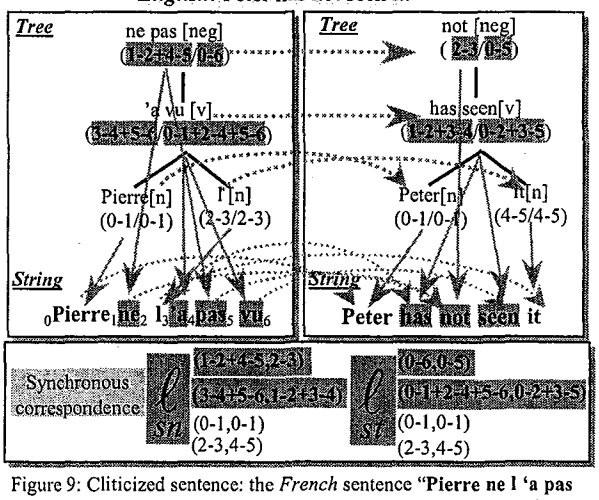

Figure 9: Cliticized sentence: the *French* sentence "**Pierre ne l 'a pas vu**" and its corresponding *English* sentence "**Peter has not seen it**".

5. DISCUSSION

As we mentioned earlier, there is now a consensus about the fact that natural language should be described as a correspondence between different levels of representation. Therefore basic linguistic structures must be created in such a way that manifests this. The proposed synchronization of two SSTCs seems to be an appropriate representation that makes such correspondence explicit. Machine translation (MT) is a promising application area for synchronous SSTC. The extended flexibility of the synchronous SSTC in recording the synchronous correspondences between the two languages enables to match linguistic units at different internal levels of their structures. This makes synchronous SSTC very well suited to be used as annotation for translation systems that automatically extract transfer mappings (rules or examples) from bilingual corpora. [13] presented an approach for constructing a Bilingual Knowledge bank (BKB) based on the synchronous SSTC, which is used by an English-Malay translation system based on the approach presented by [2]. We conclude this paper with some interesting observations on the synchronous SSTC:

i- A way to specify bi-directional structural transfers in a reasoned manner, as SSTC is used to specify structural analyzers or generators (bi-directional).

ii- Transfer rules are stated as corresponde-nces between nodes and subtrees of the trees of Synchronous SSTC associated with lexical entries. We can thus define lexical transfer rules over large domain of locality.

iii- A way to put the representation trees (i.e. a text and its translation) in a very fine-grained correspondence.

iv- The transfer between two languages, such as source and target languages in machine translation, 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 formalisms.

v- The flexibility in recording the correspondences between two languages in synchronous SSTC can be easily extended to record the correspondences between more than two languages, especially in constructing multilingual knowledge banks (MKB) (i.e. synchronization between multi languages).

vi- Synchronous SSTC inherits from the SSTC the independence from the choice of the tree structure and linguistic theories. Also the ability of handling the non-standard cases in Natural language and between different languages

REFERENCES

[1] A. Abeillé, Y. Schabes, and A. Joshi, Using lexicalized TAGs for machine translation, In Proceedings of *COLINGS'90*, Helsinki, 1990.

[2] M.H. Al-Adhaileh, and E.K. Tang, Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema. In Proceedings of MTS-VII (*Machine Translation SUMMIT VII*). Singapore, 1999.

[3] C. Boitet, and Y. Zaharin, Representation trees and string-tree correspondences. In Proceedings of COLING'88, Hungary, 1988, pp 59-64.

[4] K. Harbusch, and P. Poller, Structural Translation with Synchronous Tree Adjoining Grammars in VERBMOBILE, VERBMOBILE Report 184, Saarbrücken, Germany, 1996.

[5] K. Harbusch, and P. Poller, Non-Isomorphic Synchronous TAGs, In A. Abeillé, O. Rambow (eds). Tree Adjoining Grammars: Formal Properties, Linguistic Theory and Applications, CSLI, Stanford, California/USA, 2000.

[6] S. Kahane, What is a Natural Language and How to Describe It? Meaning-Text Approaches in Contrast with Generative Approaches, In Proceedings of 2nd International conference of Computational Linguistics and Intelligent Text Processing (*CICLing*), Mexico, 2001, pp. 1-17.

[7] I. Mel'čuk, Vers une Linguistique Sens-Texte, LeÇon inaugurale au Collège de France, Paris: Collège de France, 1997

[8] J. Milićević, A short guide to the Meaning-Text linguistic theory, in Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing (*CICLing*), Mexico, 2001.

[9] O. Rambow, and G. Satta, Synchronous Models of language In Proceedings of the 34th Meeting of the Association for Computational Linguistics (*ACL'96*), 1996.

[10] S. Shieber, Restricting the weak generative capacity of Synchronous Tree Adjoining Grammar, *Computational Intelligence*, 10(4), 1994, pp 371-385.

[11] S. Shieber, and Y. Schabes, Synchronous Tree Adjoining Grammars. In Proceedings of *COLINGS'90*, Helsinki, 1990.

[12] E. K. Tang, Natural Language Analysis in Machine Translation (MT) Based on the String-Tree Correspondence Grammar (STCG). Ph.D. thesis, Universiti Sains Malaysia, Penang, Malaysia, 1994.

[13] E.K. Tang, and M.H. Al-Adhaileh, Converting a Bilingual Dictionary into a Bilingual Knowledge Bank Based on the Synchronous SSTC Annotation Schema, In Proceedings of MTS-VIII (*Machine Translation SUMMIT VIII*), Spain, 2001.

[14] E.K. Tang, and Y. Zaharin, Handling Crossed Dependencies with the STCG. In Proceedings of *NLPRS'95*, Korea, 1995.

[15] A. Žolkovski and I. Mel'čuk, On a possible method, an instruments for semantic synthesis (of texts), *Scientific and Technological Information*, 6, 1965, pp 23-28.