A SYNCHRONIZATION STRUCTURE OF SSTC AND ITS APPLICATIONS IN MACHINE TRANSLATION

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

ZAHARIN YUSOFF
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, MALAYSIA
zarin@cs.usm.my

ABSTRACT

In this paper, a flexible annotation schema called (SSTC) is introduced. In order to describe the correspondence between different languages, we propose a variant of SSTC called synchronous SSTC (S-SSTC). We will also describe how S-SSTC provides the flexibility to treat some of the non-standard cases, which are problematic to other synchronous formalisms. The proposed S-SSTC schema is 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). Also it can be used as annotation for translation systems that automatically extract transfer mappings (rules or examples) from bilingual corpora. The S-SSTC is very well suited for the construction of a Bilingual Knowledge Bank (BKB), where the examples are kept in form of S-SSTCs.

KEYWORDS: parallel text, Structured String-Tree Correspondence (SSTC), Synchronous SSTC, Bilingual Knowledge Bank (BKB), Tree Bank Annotation Schema.

1. INTRODUCTION

There is now a consensus about the fact that natural language should be described as correspondences between different levels of representation. Much of theoretical linguistics can be formulated in a very natural manner as stating correspondences (translations) between layers of representation structures (Rambow & Satta, 1996).

In this paper, a flexible annotation schema called Structured String-Tree Correspondence (SSTC) (Boitet & Zaharin, 1988) 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 some non-standard cases (e.g. non-projective correspondence).

While synchronous systems are becoming more and more popular, there is therefore a great need for formal models of corresponding different levels of representation structures. Existing synchronous systems face a problem of handling, in a computationally attractive way, some non-standard phenomena exist between NLs. Therefore there is a need for a flexible annotation schema to realize additional power and flexibility in expressing the desired structural correspondences between languages (representation structures).

Many problems in Machine Translation (MT), in particular transfer-rules extraction, EBMT, etc., can be expressed via correspondences. We will define a variant of SSTC called synchronous SSTC (S-SSTC). S-SSTC consists of two SSTCs that are related by a synchronization relation. The use of S-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 in the case of MT. The S-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 a synchronization relation between two SSTCs, which is totally non-directional.

In this paper, we will present the proposed S-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 (Kahane, 2001). The MTT point of view, even if it has been introduced in different formulations, is more or less accepted by the whole linguistic community.

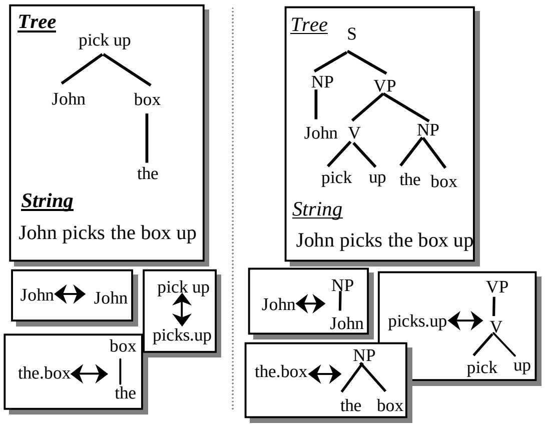


Figure 1: The correspondence between the string **"he picks the box up"** and its representation tree (dependency tree and phrase-structure tree), together with the sub-correspondences between the substrings and subtrees.

In this section, 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 linguistic formalisms; namely, the text, its corresponding abstract linguistic representation and the mapping (correspondence) between these two.

Actually, 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 sub-correspondences between parts of the string (substrings) and parts of the tree structure (subtrees), which can be interpreted for both analysis and generation in NLP. It is well known that many linguistic constructions are not projective (e.g. scrambling, cross serial dependencies, etc.). Hence, it is very much desired to define the correspondence in a way to be able to handle the non-standard cases (e.g. non-projective correspondence), see Figure 1. Towards this aim, a flexible annotation structure called Structured String-Tree Correspondence (SSTC) was introduced in Boitet & Zaharin (1988) to record the string of terms, its associated representation structure and the mapping between the two, which is expressed by the sub-correspondences recorded as part of a SSTC.

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 (Boitet & Zaharin, 1988). 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 (Tang & Zaharin, 1995).

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 root.*

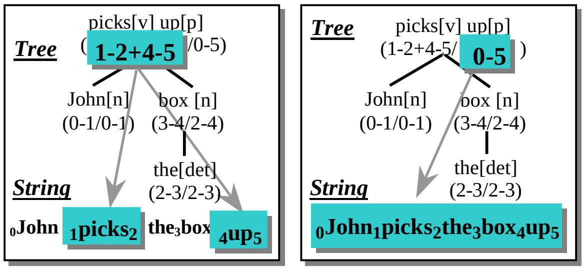


Figure 2: An SSTC recording the sentence "**John picks the box up**" and its dependency tree together with the correspondences between substrings of the sentence and subtrees of the tree.

Figure 2 illustrates the sentence **"John picks the box up"** with its corresponding SSTC. It contains a non-projective correspondence. 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

[1] The Meaning-Text Theory (MTT) was put forward in (Žolkovski & Mel'čuk (1965), in the framework of research in Machine translation. More presentations of MTT can be found in (Mel'čuk, 1997) and (Milićević, 2001).

[2] These definitions are based on the discussion in (Tang, 1994) and Boitet & Zaharin (1988).

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 correspondence between subtrees and substrings are denoted by the interval assigned to the STREE of each node, e.g. the subtree rooted at node “**picks up**” with STREE interval (0-5) corresponds to the whole sentence “**John picks the box up**”.

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"**. For more details on the proprieties of SSTC, see Boitet & Zaharin (1988).

3. SYNCHRONOUS SSTC STRUCTURE

Much of theoretical linguistics can be formulated in a very natural manner as stating correspondences (translations) between layers of representation structures (Rambow & Satta, 1996), such as the relation between syntax and semantic. An analogous problem is to be defined in such a way that expresses the correspondence between a language and its translations in other languages. Therefore the synchronization of 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. For example, synchronous Tree Adjoining Grammar (S-TAG) can be used to relate TAGs for two different languages, for example, for the purpose of immediate structural translation in machine translation (Abeillé et al.,1990), (Harbusch & Poller,1996), or for relating a syntactic TAG and semantic one for the same language (Shieber & Schabes,1990). S-TAG is a variant of Tree Adjoining Grammar (TAG) introduced by (Shieber & Schabes,1990) to characterize correspondences between tree adjoining languages. 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. It allows the construction of a non-TAL (Shieber, 1994), (Harbusch & Poller, 2000). As a result, Shieber (1994) propose a restricted definition for S-TAG, namely, the IS-TAG for isomorphic S-TAG. In this case only TAL 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 IS-TAG, Figure 3 illustrates some examples (Shieber, 1994). Similar limitations also appear in synchronous CFGs (Harbusch & Poller,1994).

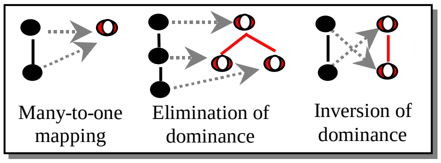


Figure 3: Kinds of relations between different languages, which are not isomorphic.

Due to these limitations, instead of investigating into the synchronization of two grammars, 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 at the level of language sentence pairs. For example, such schema can serve as a mean to represent translation examples, or find structural correspondences for the purpose of transfer grammar learning (Menezes & Richardson, 2001), (Aramaki et al., 2001), (Watanabe et al., 2000), (Meyers et al., 2000), (Matsumoto et al., 1993), (kaji et al., 1992), and example-base machine translation EBMT[3] (Sato & Nagao, 1990), (Sato, 1991), (Richardson et al., 2001), (Al-Adhaileh & Tang, 1999).

3.1 The Synchronous SSTC

In this section, we will discuss the definition and the formal properties of S-SSTC. A S-SSTC consists of a pair of SSTCs with an additional synchronization relation between them. The use of S-SSTC is motivated by the desire to describe not only the correspondence between the text and its representation structure in one language (i.e. SSTC) but also the correspondence between two languages (synchronous correspondence).

Definitions:

- *Let each of **S** and **T** be SSTC which consists of 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**, as defined in Section 2.1.*

- *A synchronous SSTC S_{syn} is defined as a triple (**S**, **T**, $\varphi_{(S,T)}$), where $\varphi_{(S,T)}$ is a set of links defining the synchronization correspondence between **S** and **T** at different internal levels of the two SSTC structures.*

- *A link $\ell \in \varphi_{(S,T)}$ can be either of type ℓ_{sn} or ℓ_{st} which defines the synchronous correspondences between nodes of **tr** in **S**, and nodes of **tr** in **T**.*

    - *ℓ_{sn} records the synchronous correspondences at level of nodes in **S** and **T** (i.e. lexical correspondences between specified nodes), and*

[3] for a comprehensive overview about EBMT, see Somers(1999)

normally ℓ_{sn} *= (X₁, X₂), where X_1 and X_2 are sequences of* *<u>SNODE</u>* *correspondences in co, which may be empty.*

- ℓ_{st} *records the synchronous correspondences at level of subtrees in **S** and **T** (i.e. structural correspondences between subtrees), and normally* ℓ_{st} *= (Y₁, Y₂), where Y_1 and Y_2 are sequences of <u>STREE</u> correspondences in co, which may be empty.*

- *A synchronous correspondence link* $\ell \in \varphi_{(S,T)}$ *can be of type* ℓ_{sn} *or* ℓ_{st} *.*

- ℓ_{sn} *is a pair(* ℓ_{sn_s} *,* ℓ_{sn_t} *), where* ℓ_{sn_s} *is from the first SSTC and* ℓ_{sn_t} *is from the second SSTC .*

- ℓ_{sn} *is represented by sets of intervals such that:*
  - ℓ_{sn_s} $= \{\ \boldsymbol{i_1\_j_1} + \ldots + \boldsymbol{i_k\_j_k} + \ldots + \boldsymbol{i_p\_j_p}\ \} \mid \boldsymbol{i_k\_j_k} \in$ ***X:SNODE*** *correspondence in co of the first SSTC.*
  - ℓ_{sn_t} $= \{\ \boldsymbol{i_1\_j_1} + \ldots + \boldsymbol{i_k\_j_k} + \ldots + \boldsymbol{i_p\_j_p}\ \} \mid \boldsymbol{i_k\_j_k} \in$ ***X:SNODE*** *correspondence in co of the second SSTC.*

- ℓ_{st} *is a pair(* ℓ_{st_s} *,* ℓ_{st_t} *), where* ℓ_{st_s} *from the first SSTC and* ℓ_{st_t} *from the second SSTC as defined below:*
  - ℓ_{st_s} $= \{\ \boldsymbol{i_1\_j_1} + \ldots + \boldsymbol{i_k\_j_k} + \ldots + \boldsymbol{i_p\_j_p}\ \} \mid \boldsymbol{i_k\_j_k}$ $\in$ ***Y:STREE*** *correspondence in co of the first SSTC or* $\boldsymbol{(i_k\_j_k) = (i_k\_j_k) - (i_u\_j_v) \mid i_u \geq i_k \wedge j_v \leq j_h :}$ **i.e.** $\boldsymbol{(i_u\_j_v) \subseteq (i_k\_j_k)}$ *which corresponds to an incomplete subtree.*
  - ℓ_{st_t} $= \{\ \boldsymbol{i_1\_j_1} + \ldots + \boldsymbol{i_k\_j_k} + \ldots + \boldsymbol{i_p\_j_p}\ \} \mid \boldsymbol{i_k\_j_k}$ $\in$ ***Y:STREE*** *correspondence in co of the second SSTC or* $\boldsymbol{(i_k\_j_k) = (i_k\_j_k) - (i_u\_j_v) \mid i_u \geq i_k \wedge j_v \leq j_h :}$ **i.e.** $\boldsymbol{(i_u\_j_v) \subseteq (i_k\_j_k)}$ *which corresponds to an incomplete subtree.*

- *The synchronous correspondence between terminal nodes with* ***<u>X:SNODE</u>*** *=* ***<u>Y:STREE</u>*** *will be of both* ℓ_{sn} *and* ℓ_{st} *correspondence such that* $\ell_{sn} = \ell_{st}$ *.*

Note: The synchronous correspondences can be between SSTCs that contain non-standard phenomena; i.e. featursiation and discontinuity (crossed dependency). In these cases the synchronous correspondence is strait forward (following the above definitions); e.g. see Figure 4 and Figure 6.

The S-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 S-SSTC is non-directional. S-SSTC is defined to make such relation explicit. Figure 4 depicts a S-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.

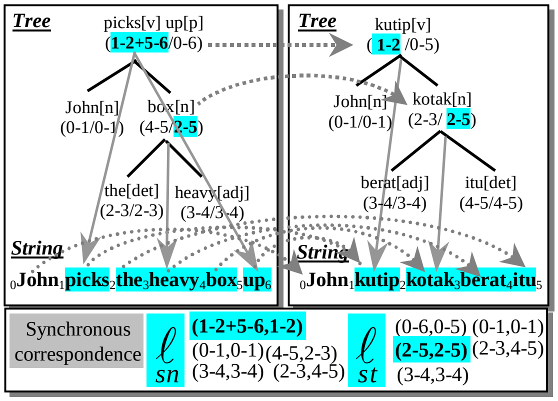


Figure 4: A synchronous SSTC for the sentence "**John picks the heavy box up**" and its Malay translation "**John kutip kotak berat itu**", together with the synchronous correspondence between them.

Based on the notation used in S-SSTC, Figure 4 illustrates the S-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 <u>SNODE</u> pairs for ℓ_{sn} and <u>STREE</u> pairs for ℓ_{st}. For ℓ_{sn} each pair is of (ℓ_{sn_s}, ℓ_{sn_t}), where ℓ_{sn_s} is <u>SNODE</u> interval/s from the *source* SSTC and ℓ_{sn_t} is <u>SNODE</u> interval/s from the *target* SSTC. As for ℓ_{st} each pair is of (ℓ_{st_s}, ℓ_{st_t}), where ℓ_{st_s} is <u>STREE</u> interval/s from the *source* SSTC and ℓ_{st_t} is <u>STREE</u> interval/s from 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 (ℓ_{sn_s}, ℓ_{sn_t})⇔(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 $(\underset{s}{\ell_{st}},\underset{t}{\ell_{st}})\Leftrightarrow$(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 $(\underset{s}{\ell_{sn}},\underset{t}{\ell_{sn}})\Leftrightarrow$(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 $(\underset{s}{\ell_{st}},\underset{t}{\ell_{st}})\Leftrightarrow$(2-5,2-5) under the ℓ_{st} synchronous correspondence.

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 (e.g. Shieber (1994) cases). Due to lack of space we will only brief on some of these non-standard cases without going into the details.

Figure 4 illustrates a case where the *English* sentence has non-standard cases of featurisation, crossed dependency and a many-to-one synchronous correspondence in “**picks up**”. 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*.

Figure 5, 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 into 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*.

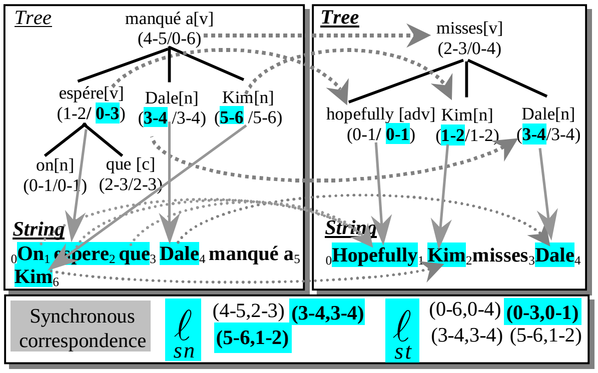


Figure 5: Many-to-one correspondence and arguments swapping correspondence in the *French* sentence “**On espére que Dale manqué a Kim**” and its corresponding *English* sentence “**Hopefully Kim misses Dale**”.

Figure 6 describes the cases of clitic climbing in French and the non-projective correspondence (i.e. crossed dependency). It shows the flexibility of SSTC and the proposed S-SSTC in handling such popular cases.

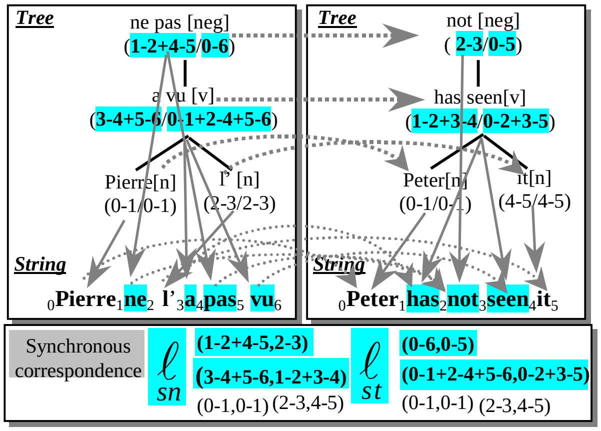


Figure 6: Cliticized sentence: the *French* sentence “**Pierre ne l ‘a pas vu**” and its corresponding *English* sentence “**Peter has not seen it**”.

Figure 7 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 (i.e. elimination of dominance). Another even more extreme relationship between the synchronized pair involving inverted correspondences is exemplified in Figure 8.

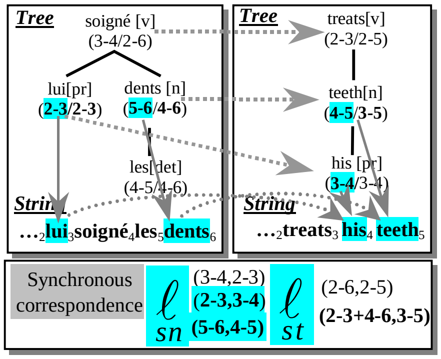


Figure 7: Elimination of dominance, in the French sentence “**le docteur lui soigné les dents**” and its corresponding English sentence “**the doctor treats his teeth**”.

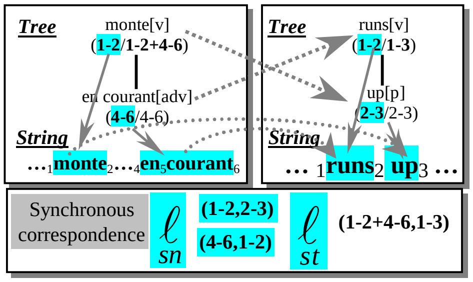


Figure 8: Inversion of dominance in the *French* sentence “**Jean monte la rue en courant**” and its corresponding *English* sentence “**John runs up the street**”.

Figure 9, depicts the case when partial subtree/s from the first SSTC has/ve a synchronous correspondence with partial subtree/s in the second SSTC. The *German* word "**beschenkte**" corresponds to the *English* phrase "**give present**" which is a partial subtree from the tree rooted by the word "**give**" in the *English* SSTC. This synchronous correspondence is recorded under the ℓ_{st} where the operation (-: minus) is used to calculate the Y:STREE interval/s for the partial subtree/s.

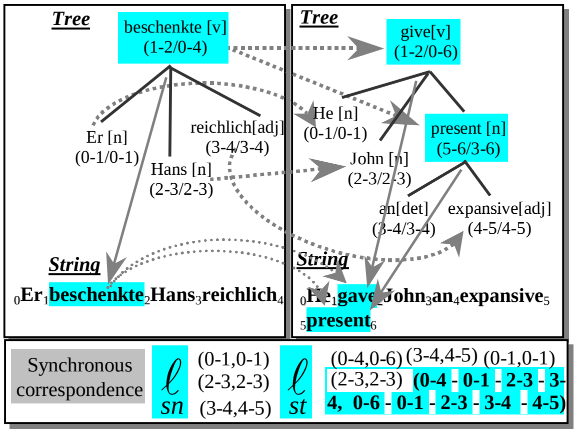


Figure 9: Partial subtree/s correspondence: the *German* sentence "**Er beschenkte Hans reichlich**" and its corresponding *English* sentence "**He gave John an expensive present**"; i.e. the use of (-) operation to calculate the Y:STREE interval.

5. SYNCHRONOUS CORRESPONDENCE CONSTRAINTS BETWEEN NATURAL LANGUAGES (NLs)

As we mentioned in Section 2, in the SSTC the correspondences between the surface text and the associated representation tree structure are ensured by means of intervals; i.e. (X:SNODE, Y:STREE). This explicitly indicates which word/s of the text correspond/s to which node in the tree. For describing a NL using SSTC, a set of constraints were defined to govern such correspondences (Lepage, 1994):

*- **X:SNODE** and **Y:STREE** intervals are governed by the following constraints:*

1. ***i) Global correspondence:** an entire tree corresponds to an entire sentence.*
2. ***ii) Inclusion:** a subtree which is part of another subtree **T**, must correspond to a substring in the substring corresponding to **T**.*
3. ***iii) Membership:** a node in a subtree **T**, must correspond to a word which is member of the substring corresponding to **T**.*

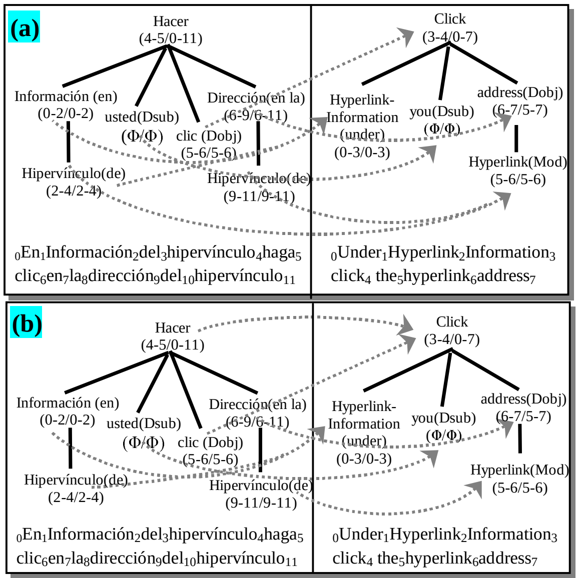
In a similar manner, in order to describe the synchronous correspondences between NLs using S-SSTC, we define a set of constraints to govern the synchronous correspondences between the different NLs. These constraints will be used to make explicitly the synchronous correspondences in a natural manner.

*- ℓ_{sn} and ℓ_{st} are governed by the following **constraints**:*

- ***Singleness:*** *A node **N** which has a synchronization correspondence, can participate in one and only one $\ell \in \ell_{sn}$, and one and only one $\ell \in \ell_{st}$. This means allowing one-to-one, one-to-many and many-to-many, but the mappings do not overlap.*
- ***Inclusion:*** *Given two ℓ_{st} correspondence pairs $\ell_{st_1} = (\ell_{st_{s_1}}, \ell_{st_{t_1}})$ and $\ell_{st_2} = (\ell_{st_{s_2}}, \ell_{st_{t_2}})$, ℓ_{st_1} and ℓ_{st_2} satisfy the inclusion constraint if and only if $\ell_{st_{s_1}} \subseteq \ell_{st_{s_2}}$ and $\ell_{st_{t_1}} \subseteq \ell_{st_{t_2}}$.*
- ***Membership:*** *Given two correspondence pairs $(\ell_{st_s}, \ell_{st_t}) \in \ell_{st}$ and $(\ell_{sn_s}, \ell_{sn_t}) \in \ell_{sn}$, ℓ_{sn} and ℓ_{st} satisfy the membership constraints if and only if $\ell_{sn_s} \subseteq \ell_{st_s}$ and $\ell_{sn_t} \subseteq \ell_{st_t}$. This means the lexical correspondences are always members in the structural correspondences.*
- ***Dominance:*** *Given two subtrees **S** and **T**, there is a correspondence $\ell \in \ell_{st}$ between **S** and **T** satisfy the dominance constraints if and only if $\forall \ell \subseteq$ **STREE(S)** correspond to $\forall \ell \subseteq$ **STREE(T)**.*
- ***Globality:*** *Given a S-SSTC, there must be $\ell \in \ell_{st}$ satisfies the globality constraints between the the root node R_s of the entire tree in the first SSTC and the root node R_t of the entire tree in the second SSTC, if and only if $(\ell_{st_s}, \ell_{st_t}) \in \ell_{st}$ such that $\ell_{st_s} =$ **STREE(R_s) : INT**(String) in the first SSTC, and $\ell_{st_t} =$ **STREE(R_t) : INT**(String) in the second SSTC. This mean the whole tree in the first SSTC corresponds to the whole tree in the second SSTC, and the whole string in the first SSTC corresponds to the whole string in the second SSTC).*

Note that these constraints can be used to license only the linguistically meaningful synchronous correspondences between the two SSTCs of the S-SSTC (i.e. between the two languages). For instance, when building translation units in EBMT approaches (Richardson et al., 2001), (Aramaki, 2001), (Al-Adhaileh &Tang, 1999), (Sato & Nagao, 1990), (Sato, 1991), (Sadler & Vendelmans, 1990), etc., where S-SSTC can be used to represent the entries of the BKB or when S-SSTC used as an annotation schema to find the translation correspondences (lexical and structural correspondences) for transfer-rules' extraction from parallel parsed corpus (Menezes & Richardson, 2001), (Watanabe et al., 2000), (Meyers et al., 2000), (Matsumoto et al., 1993) and (kaji et al., 1992). Note that the grammar alignment rules used in (Menezes & Richardson, 2001) can be reformulated using these constraints to construct the transfer mappings from a synchronous *source-target* example.

Figure 10 shows an example from Menezes and Richardson (2001), the logical form for the Spanish-English pair: ("***En Información del hipervínculo, haga clic en la dirección del hipervínculo***", "***Under Hyperlink Information, click the hyperlink address***").



Synchronous correspondence	ℓ_{sn}	(5-6,3-4) (6-9,6-7) (Φ, Φ) (9-11,5-6)	ℓ_{st}	(0-11,0-7) (0-4,0-3) (Φ, Φ) (6-11,5-7) (9-11,5-6)

Figure 10: (a) the lexical correspondences, (b) the structural correspondences after applying the constraints.

Recently, the development of machine translation systems requires a substantial amount of translation knowledge typically embodied in the bilingual corpora. For instance, the development of translation systems based on transfer mappings (rules or examples) that automatically extracted from these bilingual corpora. All these systems typically first obtain a tree structures (normally a predicate-argument or a dependency structure) for both the source and target sentences. From the resulting structures, lexical and structural correspondences between the two structures are extracted, which are then presented as a set of examples in a bilingual knowledge bank (BKB) or transfer rules for translation process.

However, what has so far been lacking is a schema or a framework to annotate and express such extracted lexical and structural correspondences in a flexible and powerful manner. The proposed S-SSTC annotation schema can fulfill this need, and it is flexible enough to handle different type of relations that may happen between different languages' structures. S-SSTC very well suited for the construction of a BKB, which is needed for the EBMT applications. Al-Adhaileh and Tang (2001) presented an approach for constructing a BKB based on the S-SSTC.

In S-SSTC, the synchronous correspondence is defined in a way to ensure a flexible representation for both lexical and structural correspondences: ***i-*** Node–to–node correspondence (lexical correspondence), which is recorded in terms of pair of intervals ($\boldsymbol{X_s}$,$\boldsymbol{X_t}$) where $\boldsymbol{X_s}$ and $\boldsymbol{X_t}$ is SNODE interval/s for the source and the target SSTC respectively, ***ii-*** Subtree–to–Subtree correspondence (structural correspondence), which is very much needed for relating the two different languages at a level higher than the lexical level, a level of phrases. It is recorded in terms of pair of intervals ($\boldsymbol{Y_s}$,$\boldsymbol{Y_t}$) where $\boldsymbol{Y_s}$ and $\boldsymbol{Y_t}$ is STREE interval/s for the source and the target SSTC respectively.

Furthermore, the SSTC structure can easily be extended to keep multiple levels of linguistic information, if they are considered important to enhance the performance of the machine translation system (i.e. ***Features transfer***). For instance, each node representing a word in the annotated tree structure can be tagged with part of speech (POS), semantic features and morphological features.

6. CONCLUSION

The proposed S-SSTC is not limited for the case discussed here (i.e. MT), any system need to describe two language structures and the synchronization relation between them, can used S-SSTC as annotation schema for that. This is for example the case for presenting the syntax-semantics interface between different languages. S-SSTC is a flexible schema, which is able to handle non-standard phenomena that may occur between different languages. We conclude this paper with some interesting observations on the synchronous SSTC:

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

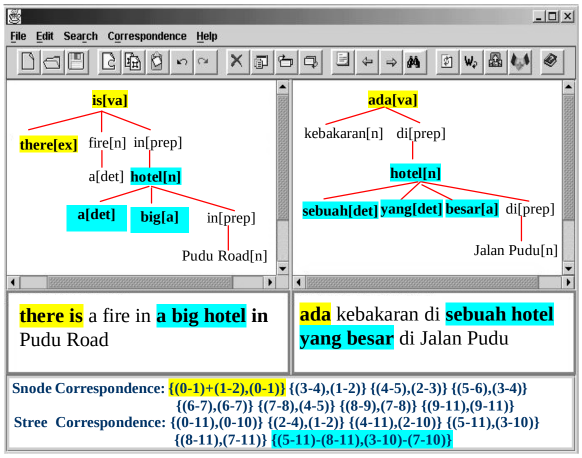
ii- A natural way to specify bi-directional structural transfer, as SSTC is used to specify structural analyzers and generators (i.e. bi-directional).

iii- Synchronous SSTC can be easily extended to record the correspondences between more than two languages, hopefully with transitive property, especially in constructing multilingual knowledge banks (MKB) (i.e. synchronization between multiple languages).

iv- 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.

v- 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.

Also a GUI editor has been implemented for view, edit, create and correct the S-SSTC components, as illustrated in Figure 11.



there is a fire in **a big hotel in** Pudu Road

ada kebakaran di **sebuah hotel yang besar** di Jalan Pudu

Snode Correspondence: {(0-1)+(1-2),(0-1)} {(3-4),(1-2)} {(4-5),(2-3)} {(5-6),(3-4)} {(6-7),(6-7)} {(7-8),(4-5)} {(8-9),(7-8)} {(9-11),(9-11)}

Stree Correspondence: {(0-11),(0-10)} {(2-4),(1-2)} {(4-11),(2-10)} {(5-11),(3-10)} {(8-11),(7-11)} {(5-11)-(8-11),(3-10)-(7-10)}

Figure 11: Synchronous SSTC Editor.

REFERENCES

Abeillé, A., Schabes, Y. and Joshi, A. (1990). Using lexicalized TAGs for machine translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLINGS'90)*, Helsinki, Finland, pp 1-6.

Al-Adhaileh, M.H. and Tang, E.K. (1999). Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema. In *Proceedings of Machine Translation Summit VII*, Singapore, pp 244-249.

Al-Adhaileh, M.H. and Tang, E.K. (2001). Converting a Bilingual Dictionary into a Bilingual Knowledge Bank Based on the Synchronous SSTC Annotation Schema. In *Proceedings of Machine Translation Summit VIII*. Spain, pp 351-356.

Boitet, C. and Zaharin, Y. (1988). Representation trees and string-tree correspondences. In *Proceedings of the 12th International Conference on Computational Linguistics (COLINGS-88)*, Budapest. Hungary, August, pp 59-64.

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

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

Kaji H., Kida Y., and Morimoto Y. (1992), Learning Translation Templates from Bilingual Text. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, France, pp 672-678.

Lepage, Y. (1994). Texts and Structures – Pattern-matching and Distances, ATR report TR-IT-0049, Kyoto, Japan.

Matsumoto, Y., Ishimoto H., and Utsuro, T. (1993). Structural Matching of Parallel Texts. In *Proceedings of the 31th annual meeting of Association for Computational Linguistics (ACL-93)*, pp 23-30.

Mel'čuk I. (1997). *Vers une Linguistique Sens-Texte*. Leçon inaugurale au Collège de France, Paris: Collège de France.

Menezes, A. and Richardson, S. (2001). A Best-first Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In *the workshop on Data-Driven Machine Translation, at the 38th Annual Meeting of the Association for Computational Linguistic (ACL 2001)*, Toulouse, France.

Meyers A., Kosaka M., and Grishman R. (2000). Chart-based Transfer Rule Application in Machine Translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, Germany / Luxembourg.

Meyers, A., Yangarber, R. and Grishman R. (1996). Alignment of Shared Forests for Bilingual Corpora. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, pp 459-465.

Milićević J. (2001). A short guide to the Meaning-Text linguistic theory. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, Coleccion en Ciencias de Computacion, Fondo de Cultura Economica- IPN - UNAM*, Mexico.

Rambow, O. and Satta, G. (1996). Synchronous Models of language. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96)*, Santa Cruz, CA, USA.

Richardson S., Dolan W., Menezes A. and Pinkham J. (2001). Achieving commercial-quality translation with example-based methods. In *Proceedings of Machine Translation SUMMIT VIII*, Spain, pp 293-297.

Sadler V. and Vendelmans, R. (1990). Pilot Implementation of a Bilingual Knowledge Bank. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Vol. 3, Helsinki, Finland, pp 449-451.

Sato S. and Nagao M. (1990). Towards Memory-based Translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, Vol. 3, Helsinki, Fenland, pp 247-252.

Sato, S. (1991). *Example-Based Machine Translation*. Ph.D. thesis, Kyoto University, Japan.

Shieber, S. (1994). Restricting the Weak Generative Capacity of Synchronous Tree Adjoining Grammar. *Computational Intelligence*, 10(4): 371-385.

Shieber, S. and Schabes, Y. (1990). Synchronous Tree Adjoining Grammars. In *Proceedings of the 13th International Conference on Computational Linguistics (COLINGS-90)*, Helsinki, Finland, pp 253-258.

Somers, H. (1999). Review article: Example-based Machine Translation, *Machine Translation*, 14: 113-157.

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

Tang, E. K. and Zaharin, Y. (1995). Handling Crossed Dependencies with the STCG. *In Proceedings of NLPRS'95*, Seoul, Korea.

Watanabe H., Kurohashi S., and Aramaki E. (2000). Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Luxembourg/Saarbruecken, Germany.

Žolkovski, A. and Mel'čuk, I. (1965). On a Possible Method an Instruments for Semantic Synthesis (of texts). *Scientific and Technological Information*, 6: 23-28.