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.

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 1 illustrates the sentence “**John picks the box up**” and its dependency tree together with the correspondences between the sub-strings and subtrees.

![Figure 1](image1.png)

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](image2.png)

---

\(^1\) The Meaning-Text Theory (MTT) was put forward in (Žižlava & Melčuk, 1965), in the framework of research in Machine translation. More presentations of MTT can be found in (Melčuk, 1997) and (Miličević, 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 STSC structure treats some non-standard linguistic phenomena. The particle "up" is featured 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).

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 EBMT3 (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 ($S$, $T$, $\Phi_{S,T}$), where $\Phi_{S,T}$ is a set of links defining the synchronization correspondence between S and T at different internal levels of the two SSTC structures.
- A link $\ell \in \Phi_{S,T}$ can be either of type $\ell^{st}$ or $\ell^{sn}$ which defines the synchronous correspondences between nodes of tr in $S$ and nodes of tr in $T$.
  - $\ell^{st}$ records the synchronous correspondences at $S$ and $T$ (i.e. lexical correspondences between specified nodes), and

3 for a comprehensive overview about EBMT, see Somers(1999)
normally $\ell = (X_1, X_2)$, where $X_1$ and $X_2$ are sequences of S-NODE correspondences in $\sigma$, which may be empty.

- $\ell$ records the synchronous correspondences at st level of subtrees in $\mathcal{S}$ and $\mathcal{T}$ (i.e. structural correspondences between subtrees), and normally $\ell = (Y_1, Y_2)$, where $Y_1$ and $Y_2$ are sequences of st STREE correspondences in $\sigma$, which may be empty.

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

- $\ell$ is a pair $(\ell_{sn}, \ell_{st})$, where $\ell_{sn}$ is from the first SSTC and $\ell_{st}$ is from the second SSTC.

- $\ell$ is represented by sets of intervals such that:
  - $\ell_{sn} = \{(i_1, j_1) + \ldots + (i_k, j_k) + \ldots + (i_p, j_p) \mid i_k, j_k \in \mathcal{S}:\text{S-NODE correspondence in } \sigma \text{ of the first SSTC,}\}
  - $\ell_{st} = \{(i_1, j_1) + \ldots + (i_k, j_k) + \ldots + (i_p, j_p) \mid i_k, j_k \in \mathcal{T}:\text{S-NODE correspondence in } \sigma \text{ of the second SSTC,}\}

- $\ell$ is a pair $(\ell_{sn}, \ell_{st})$, where $\ell_{sn}$ is from the first SSTC and $\ell_{st}$ is from the second SSTC as defined below:
  - $\ell_{sn} = \{(i_1, j_1) + \ldots + (i_k, j_k) + \ldots + (i_p, j_p) \mid (i_k, j_k) \in \mathcal{S}:\text{STREE correspondence in } \sigma \text{ of the first SSTC or } (i_k, j_k) = (i_j, j_i) \text{ where } i_j \geq i_k \wedge j_j \leq j_i : i.e. (i_k, j_k) \subseteq (i_j, j_i) \text{ which corresponds to an incomplete subtree,}\}
  - $\ell_{st} = \{(i_1, j_1) + \ldots + (i_k, j_k) + \ldots + (i_p, j_p) \mid (i_k, j_k) \in \mathcal{T}:\text{STREE correspondence in } \sigma \text{ of the second SSTC or } (i_k, j_k) = (i_j, j_i) \text{ where } i_j \geq i_k \wedge j_j \leq j_i : i.e. (i_k, j_k) \subseteq (i_j, j_i) \text{ which corresponds to an incomplete subtree,}\}

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

Note: The synchronous correspondences can be between SSTCs that contain non-standard phenomena; i.e. featurisation 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.

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 SNODE pairs for $\ell_{sn}$ and STREE pairs for $\ell_{st}$. For $\ell_{sn}$ each pair is of $(\ell_{sn}, \ell_{sn})$, where $\ell_{sn}$ is SNODE interval/s from the source SSTC and $\ell_{st}$ is SNODE interval/s from the target SSTC. As for $\ell_{st}$ each pair is of $(\ell_{sn}, \ell_{st})$, where $\ell_{sn}$ is STREE interval/s from the source SSTC and $\ell_{st}$ is STREE 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 $(\ell_{sn}, \ell_{st}) \Rightarrow (1-2+5-6,1-2)$ under the $\ell_{sn}$ synchronous correspondence. Whereas, the fact that “John picks the heavy box up” corresponds to “John kutip kotak berat itu”
Hello, this is a sample text as per your request.
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:
  i) **Global correspondence**: an entire tree corresponds to an entire sentence.
  ii) **Inclusion**: a subtree which is part of another subtree \( \mathcal{T} \) must correspond to a substring in the substring corresponding to \( \mathcal{T} \)
  iii) **Membership**: a node in a subtree \( \mathcal{T} \) must correspond to a word which is member of the substring corresponding to \( \mathcal{T} \)

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.

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 EBM1 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 transfers’ extraction from parallel parsed corpus (Menezes & Richardson, 2001), (Watanabe et al.,...
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”).

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

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 \((X_s, X_t)\) where \(X_s\) and \(X_t\) 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 \((Y_s, Y_t)\) where \(Y_s\) and \(Y_t\) 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.

Figure 11: Synchronous SSTC Editor.

REFERENCES


