Graphical Annotation for Syntax-Semantics Mapping

Kôiti Hasida
Social ICT Research Center,
Graduate School of Information Science and Technology,
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
E-mail: hasida.koiti@i.u-tokyo.ac.jp

Abstract

A potential work item (PWI) for ISO standard (MAP) about linguistic annotation concerning syntax-semantics mapping is discussed. MAP is a framework for graphical linguistic annotation to specify a mapping (set of combinations) between possible syntactic and semantic structures of the annotated linguistic data. Just like a UML diagram, a MAP diagram is formal, in the sense that it accurately specifies such a mapping. MAP provides a diagrammatic sort of concrete syntax for linguistic annotation far easier to understand than textual concrete syntax such as in XML, so that it could better facilitate collaborations among people involved in research, standardization, and practical use of linguistic data. MAP deals with syntactic structures including dependencies, coordinations, ellipses, transsentential constructions, and so on. Semantic structures treated by MAP are argument structures, scopes, coreferences, anaphora, discourse relations, dialogue acts, and so forth. In order to simplify explicit annotations, MAP allows partial descriptions, and assumes a few general rules on correspondence between syntactic and semantic compositions.

Keywords: diagrammatic annotation, syntax-semantics mapping, standardization

1. Introduction

A potential work item (PWI) for ISO standard (let us call it 'MAP' for convenience in the rest of the paper) of linguistic annotation concerning syntax-semantics mapping is introduced, which is an extension of SemAF-DS (ISO, 2013), which in turn is based on Linguistic DS (Description Scheme) in ISO/IEC (2004). Importing more from Linguistic DS, MAP extends some standards devised by ISO/TC37/SC4, including LAF (Linguistic Annotation Framework; ISO 2010), SynAF (Syntactic Annotation Framework; ISO 2012a), and SemAF (Semantic Annotation Framework; ISO 2012b, 2012c, 2013), while incorporating insights from relevant literature (Asher & Lascarides 2003; Carlson, et al., 2003; Haji, et al. 2006; Mann & Thompson 1988; Palmer, et al. 2005; Prasad, et al. 2008; PTB).

MAP defines how to diagrammatically annotate linguistic data to specify a mapping between its possible syntactic and semantic structures. The syntactic structures may be dependencies, coordinations, ellipses, and so forth, encompassing both intrasentential transsentential constructions, and so forth. The semantic structures consist of argument structures, scopes (of quantifications, negations, modal operators, etc.), coreferences, anaphora, and so on.

A major purpose of MAP is to facilitate collaborations among people involved in research, standardization, and practical use of linguistic annotation. For that sake, MAP provides a diagrammatic sort of concrete syntax (ISO, 2012b, 2012c) for linguistic annotations far easier to understand than traditional textual concrete syntax such as in XML. Besides being diagrammatic and intuitive, MAP is formal in the same sense that UML is formal. Namely, a MAP diagram accurately specifies a mapping between syntactic structures and semantic structures of the annotated linguistic data in question.

The rest of the paper is organized as follows. Section 2 introduces MAP diagrams to represent annotated linguistic data. Section 3 and 4 discuss further details of annotations concerning local and nonlocal compositions, respectively. Section 5 concludes the paper.

2. Annotated Segment

Let us refer to markable (annotatable) linguistic data as **segments**. A segment may be text, audio, video, etc., and may be intrasentential or transsentential. In MAP, a segment may accompany a syntactic annotation, a semantic structure, or both. Such a possibly annotated segment is diagrammatically represented by a possibly multi-part box as in Figure 1.

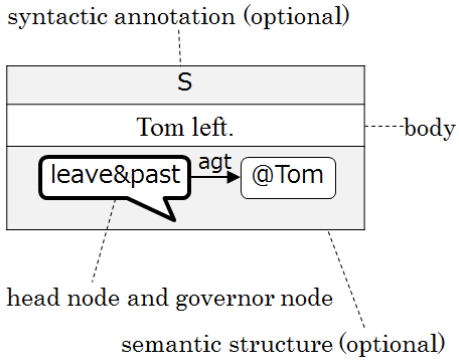

Figure 1: Annotated Segment as MAP Diagram

The top gray part of the box contains a syntactic annotation to the segment. The middle white part is the body of the whole box and contains the segment itself. As discussed later, this body part may recursively embed smaller annotated segments and, together with the syntactic-annotation part, partially specifies the syntactic structure of the segment. The bottom gray part contains a possible semantic structure of the segment. This paper assumes that semantic structures are labelled directed graphs (such as

semantic network and RDF graph) as in Figure 1, but MAP allows any other format for representing semantic structures.
Such an annotated segment defines a mapping between possible syntactic structures and possible semantic structures of the segment. The example in Figure 1 involves no syntactic ambiguity, but some examples in the rest of the paper are syntactically ambiguous so that they accommodate multiple possible syntactic structures and therefore multiple possible semantic structures.

3. Local Compositions

The semantic structure (as a labelled directed graph) annotating a segment as in Figure 1 has two designated nodes: the **head node** and the **governor node** of the segment. The head node has thick border, and the governor node is depicted as a balloon. So the leave&past node in Figure 1 is both the head node and the governor node of segment 'Tom left.' In Figure 2, the @Tom node is the head node and the empty node is the governor node of segment 'Tom.'.

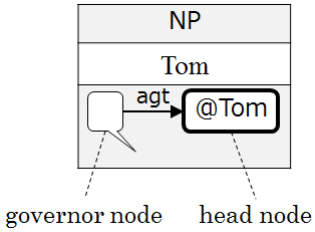

Figure 2: 'Tom' Referencing the Agent of an Action

This annotated segment represents 'Tom' as a noun phrase referring to Tom as the agent of some action represented by the governor node. In general, the governor node of segment *X* is equal to the head node segment *Y* when *X* (syntactically and hence semantically) depends on (i.e., is governed by) *Y*, as explained later.
Annotated segments may be embedded in the body part of a larger segment composed of them. There is an order among the embedded segments: from left to right and from top to bottom in the case of western languages. For instance, shown in Figure 3 is an annotated segment 'Tom left' whose body part embeds two daughter segments for 'Tom' and 'left.'

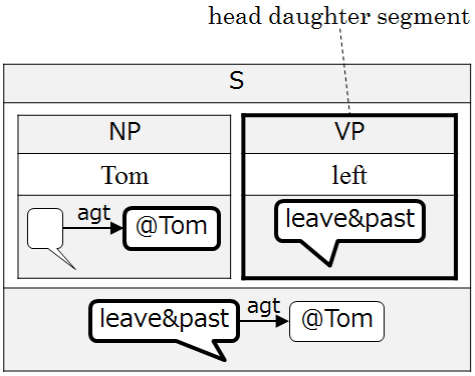

Figure 3: Local Dependency

In general, a thick-bordered daughter segment is the head daughter of the mother segment. So segment 'left' is the head of 'Tom left' in this example.
General rules in MAP for dependency constructions follow:

[1] The semantic structure of the mother segment is the union of the semantic structures of the daughter segments.
[2] The mother segment and the head daughter segments share the same head node and the same governor node.
[3] The governor nodes of the dependent daughter segments are the head node of the mother segment (which is same as the head node of the head daughter segment, due to [2]).

These rules simplify annotations. For instance, the annotated segment in Figure 3 is equivalent to the one below, because the semantic structure of the whole segment in Figure 3 is derived from those of the daughter segments by the above rules.

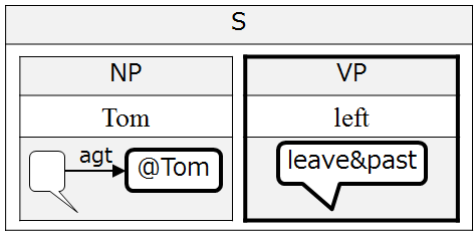

Figure 5: Simplified Annotation Equivalent to Figure 3

This is a typical annotated segment based on MAP, where only the lexical-entry segments are explicitly annotated with semantic structures and the semantic structures of larger segments are implicitly derived by the above rules.

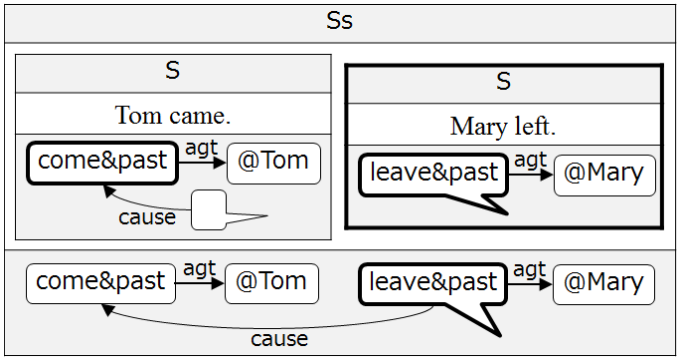

Figure 4: Intersentential Dependency

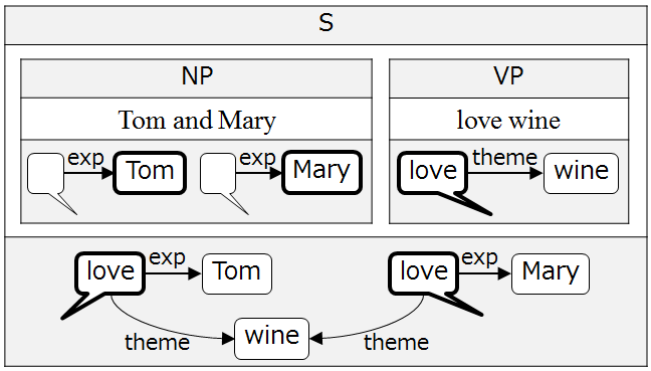

Figure 6: Semantic-Structure Duplication Due to a Distributive Coordination

The same rules apply to dependencies outside of sentences (i.e., dependencies among sentences, paragraphs, sections, and so forth), too, as follows.

Distributive coordinations are accounted for just by the abovet rule [1]. Figure 7 shows how this works, where again the semantic structure of the mother segment may be omitted thanks to the rule.

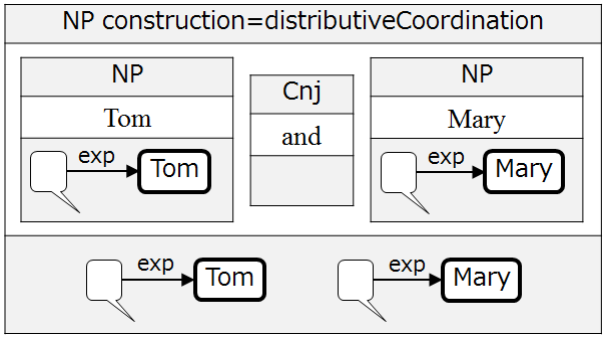

Figure 7: Distributive Coordination

This whole noun phrase and a verb phrase compose a sentence while duplicating the head node of the verb phrase as follows.

On the other hand, a collective coordination has a single head node and a single governor node, though further details are omitted in this abstract.

4. Nonlocal Compositions

MAP uses typed links to express relationships among unadjacent segments. For instance, a dep link addresses an unadjacent dependency, such as in the extraposition below.

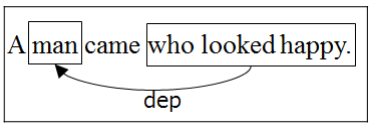

Figure 8: Extraposition

Hereafter the syntactic annotation parts and the semantic structure parts of the segments are omitted for the sake of simplicity.

An eq link addresses a coreference, as below.

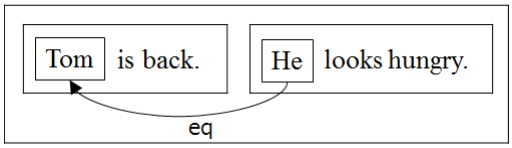

Figure 9: Coreference

Precisely speaking, an eq link represents the coreference between the head nodes of the two linked segments. So an eq link is used also for a relativization to address the coreference between the head noun and the gap in the relative clause, as follows.

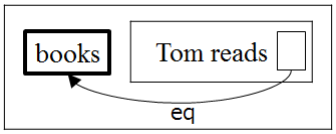

Figure 10: Relativization

A partOf link means that the head node of the source segment refers to a part of the referent of the head node of the destination segment. Below is an example of an indirect anaphora, where the parOf link means that the door is a part of the house.

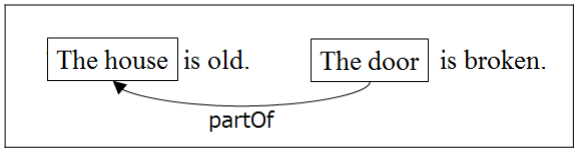

Figure 11: Indirect Anaphora

A coScope link means that the head nodes of the two linked segments belong to the same scope (of quantification, negation, modal operator, or other type of abstraction). For instance, the following example means that there is a specific woman whom every man loves, because the woman belongs to the same scope to which the state of affairs referenced by the entire sentence belongs.

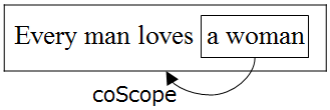

Figure 12: Wide-Scope Reading of ‘a woman’

On the other hand, the below means that different men may love different women.

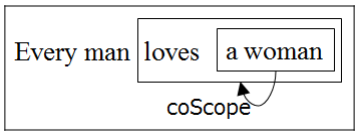

Figure 13: Narrow-Scope Reading of ‘a woman’

Similarly, in Figure 14 there is a specific doctor who Jane wants to marry, as the coScope link points to the topmost scope encompassing the entire discourse, whereas in Figure 15 there is no such specific doctor, as the coScope link there means that the marrying event and the doctor belong to the same scope of the modal operator corresponding to ‘wants.’

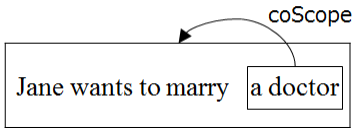

Figure 14: Wide-Scope Reading of ‘a doctor’

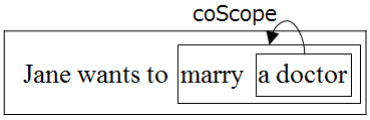

Figure 15: Narrow-Scope Reading of ‘a doctor’

The cp and subst links address ellipses, which is a reformulation of part of the Penn TreeBank (PTB) annotation scheme. For instance, the below example means that Bill wants to date with Sue, because the latter half of the sentence is interpreted by copying the former half while substituting ‘Tom’ with ‘Bill’ and ‘Mary’ with ‘Sue.’

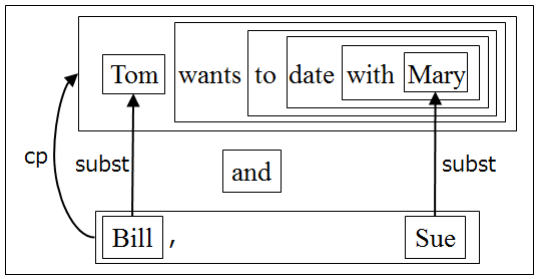

Figure 16: Ellipsis

Similarly, the below example illustrates a comparative construction involving an ellipsis, where ‘Sue’ is interpreted as ‘Tom loves Sue’ by copying ‘Tom loves Mary’ while substituting ‘Mary’ with ‘Sue.’

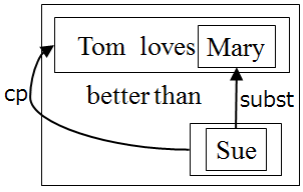

Figure 18: Ellipsis in Comparative

For instance, ‘Tom loves his wife. So does Bill.’ is ambiguous as to whether Bill loves Tom’s wife (so called strict identity) or Bill’s wife (sloppy identity).

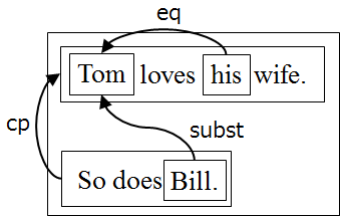

Figure 19: Ambiguity Concerning Strict/Sloppy Identity

This ambiguity is resolved by coScope links. If ‘his’ has a wider scope than ‘Tom loves his wife.’ then the copy operation excludes ‘his’ and hence the eq link as well, to infer that Bill loves Tom’s wife.

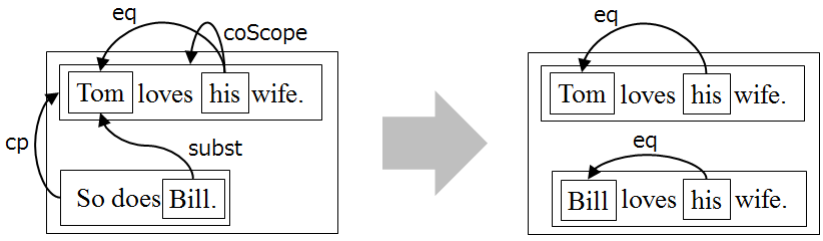

Figure 17: Strict Identity

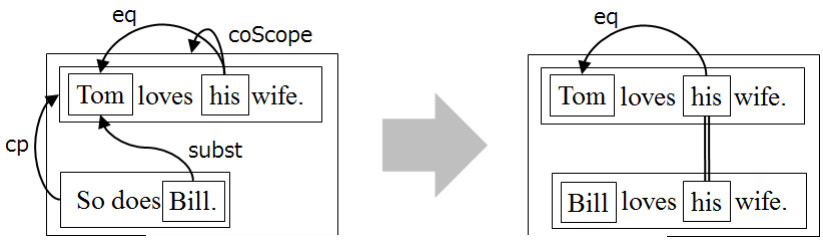

Figure 20: Sloppy Identity

If ‘his’ and ‘Tom loves his wife.’ have the same scope, on the other hand, then the copy operation involves the eq link and its destination (‘Tom’) is substituted by ‘Bill,’ which means that Bill loves Bill’s wife.

5. Final Remarks

MAP provides a diagrammatic annotation scheme to specify mappings between syntactic and semantic structures of annotated segments. In typical annotations, only the lexical-entry segments are explicitly annotated with semantic structures, and rules [1] through [3] and links among segments derive the semantic structures of larger segments.

MAP, NAF (Fokkens, et al., 2014), and NKF (NLP Annotation Knowledge-Base Format) are closely related potential work items in ISO/TC37/SC4/WG5. Since they have similar objectives and hence many common features, their relationship must be sorted out to define how to coordinate them.

References

N. Asher & A. Lascarides (2003) *Logics of Conversation*. Cambridge University Press.

L. Carlson, D. Marcu, M. E. Okurowski (2003) Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. J. van Kuppevelt & R. Smith (eds.) *Current Directions in Discourse and Dialogue*, 85-112, Kluwer Academic Publishers.

A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. van Hage, and P. Vossen (2014) NAF and GAF: Linking Linguistic Annotations. *Proceedings of 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Reykjavik, Iceland.

J. Haji, et al. (2006) Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia.

ISO (2010) ISO 24615:2010, Language resource management. Syntactic annotation framework (SynAF).

ISO (2012a) ISO 24612:2012, Language resource management. Linguistic annotation framework (LAF).

ISO (2012b) ISO 24617.1:2012, Language resource management. Semantic annotation framework . Part 1: Time and events (SemAF-Time, ISO-TimeML).

ISO (2012c) ISO 24617.2:2012, Language resource management. Semantic annotation framework . Part 2: Dialogue Acts.

ISO (2013) ISO TS 24617-5: Language Resource Management, Semantic Annotation Framework (SemAF), Part 5: Discourse structure (SemAF-DS).

ISO/IEC (2004) ISO/IEC 15938.5:2003/Amd.1:2004, Information technology. Multimedia content description interface. Part 5: Multimedia description schemes AMENDMENT 1: Multimedia description schemes extensions (MPEG-7 MDS AMD1).

W. Mann & S. Thompson (1988) Rhetorical Structure Theory: A Theory of Text Organisation. *Text*, 8(3) 243.281.

M. Palmer, D. Gildea, P. Kingsbury (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71-105.

R. Prasad. N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, et al. (2008) The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

PTB. The Penn Treebank Project. http://www.cis.upenn.edu/~treebank/