Dependency treebanks: methods, annotation schemes and tools

Introduction

- The most commonly used argument for selecting the dependency (dep.) format for building a treebank (tb) is that the tb is being created for a language with a relatively free word order (e.g. Basque, Czech, and Turkish).
- On the other hand, dep. tb have been developed to languages such as English, which have been usually seen as languages which can be better represented in the constituent formalism. The motivations for using dep. annotation vary from the fact that the level of structure is the one needed by many if not most applications to the fact that the level of representation offers a proper interface between syntactic and semantic representation
- Furthermore, dep. structures can be automatically converted in to phrase structures if needed, although usually not with 100% accuracy.
- Most of the dependency treebanks consist of written text; there is only one that is based on spoken utterances.

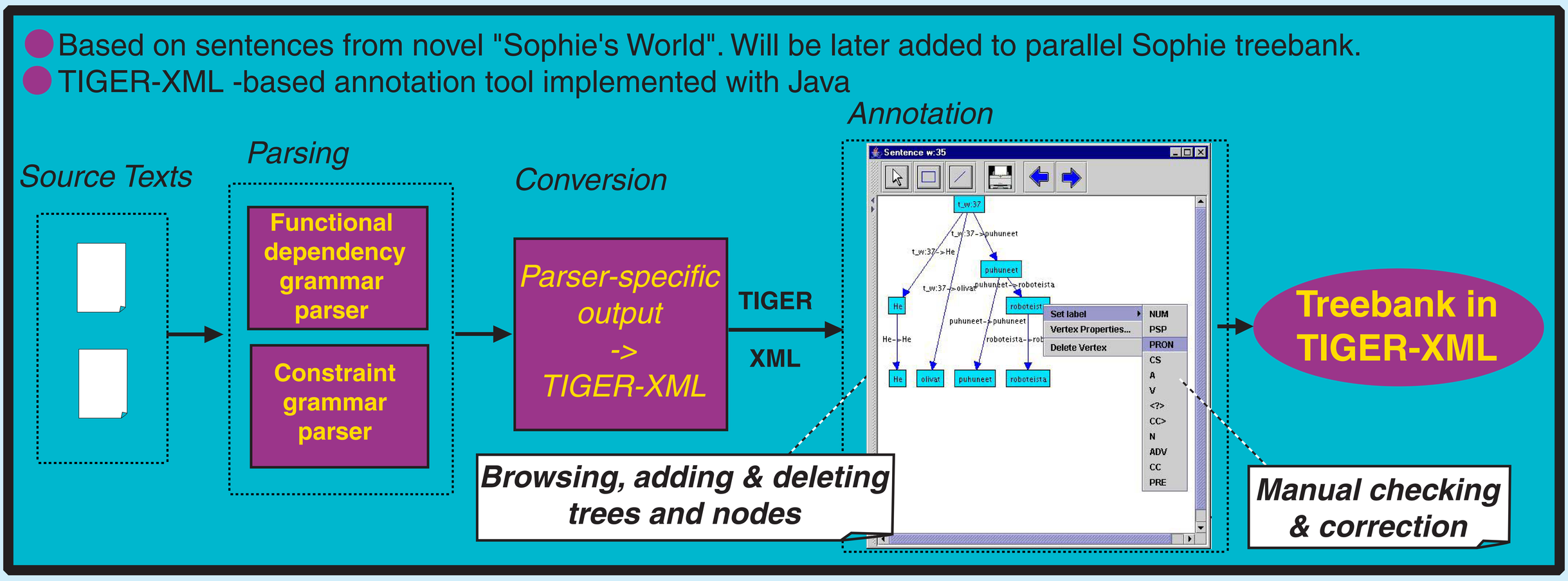
Some Dependency Treebank Projects

- There exists some tbs that have been annotated completely manually, but with taggers and parsers available to automate some of the work, such way of building tbs is rarely employed in state-of-the-art treebanking.
- The most efficient method is to perform POS and morphological tagging and at least some part of the syntactic parsing automatically, and the resulting structures are checked and corrected by human annotators.

Name	La.	Annotation	Genre	Size (sentences)	Annotation Methods	Parser	Annotation tool / Supported formats
Prague Dependency TB 1.0	Czech	23 rel. Also on the level of meaning	Newsp. (general, economic)	90000	M/SA	Lexicalized stochastic parser (Collins)	FS, CSTS SGML, Annotation Graphs XML
Danish Dependency TB	Danish	Discontinuous Grammar	Range of topics & genres	~5500	M. Morphosyn. annotation obtained from a corpus	-	PAROLE-DK format with additions, supports TIGER-XML
Alpino	Dutch	Constituent & dep.	Newsp. For parser evaluation	6000	SA, partially man. disambiguation aided by parse selection tool	HPSG-based Alpino parser	Own XML-based
Dependency TB for Russian	Russian	78 rel.	Fiction, newsp. & scientific	12000	SA	Morphol. analyzer and a parser	XML-based TEI-compatibe
TIGER TB	German	Constituent & dep.	Newsp.	50000	SA	Probabilistic parser / LFG parser	TIGER-XML
Dependency TB of English	English	Lexical predicate-argument struc.	Spoken, travel agent dial.	13000 words	SA, M correction of parser output & automatic checking of inconsistencies	Supertagger and Lightweight Dependency Analyzer	FS

Constructing A Dependency Treebank of Finnish

- Based on sentences from novel "Sophie's World". Will be later added to parallel Sophie treebank.
- TIGER-XML -based annotation tool implemented with Java



Tuomo Kakkonen
Department of Computer Science
University of Joensuu, Finland

This work was partially conducted while the author was working as a Research Fellow at MULTILINGUA, Marie Curie Early Stage Research Training Site at the University of Bergen, Norway