Lingua 120 (2010) 1567–1578

Contents lists available at ScienceDirect

Lingua

journal homepage: www.elsevier.com/locate/lingua

Dependency direction as a means of word-order typology: A method based on dependency treebanks

Haitao Liu

Institute of Applied Linguistics, Communication University of China, CN-100024, Beijing, PR China

ARTICLE INFO

Article history:
Received 14 February 2008
Received in revised form 23 April 2009
Accepted 1 October 2009
Available online 31 October 2009

Keywords:
Word-order typology
Dependency grammar
Treebank
Head-initial
Head-final

ABSTRACT

Word-order typology often uses the linear order of binary grammatical pairs in sentences to classify a language. The present paper proposes a method based on dependency treebanks as a typological means. This paper investigates 20 languages using treebanks with different sizes from 16 K to 1 million dependencies. The results show that some languages are more head-initial or head-final than others, but all contain head-initial and head-final elements. The 20 languages can be arranged on a continuum with complete head-initial and head-final patterns as the two ends. Some data about subject–verb, object–verb and adjective–noun are extracted from the treebanks for comparison with the typological studies based on the traditional means, the results are similar. The investigation demonstrates that the proposed method is valid for positioning a language in the typological continuum and the resources from computational linguistics can also be used in language typology.

1. Introduction

In typological studies on word order, the linear order of the grammatical units in a sentence is often used as the primary way to distinguish one language from another.

Greenberg (1963) is generally considered as the initiator of this field.[1] Greenberg proposed 45 linguistic universals. Twenty-eight of these universals touch on the order or position of grammatical units, for instance, the basic order of subject, object, and verb. Dryer (1992) reports the results of detailed word-order correlations based on a sample of 625 languages. Dryer defined correlation pairs as the following: "If a pair of elements X and Y is such that X tends to precede Y significantly more often in VO languages than in OV languages, then (X, Y) is a CORRELATIONS PAIR, and X is a VERB PATTERNER and Y an OBJECT PARTNER with respect to this pair" (Dryer, 1992:87). According to his investigation, there are 17 correlation pairs and 5 non-correlation pairs with the verb and object (1992:108). Dryer (1997) argues that a more useful typology is one that is based on the two binary parameters OV vs. VO and SV vs. VS. These studies demonstrate that the linear order and binary relation of two grammatical units in a sentence is an important means to catch the typological features of human languages.

It is noteworthy that although typologists examine the basic word order of a language with the possible orders of SVO, such a trigram relation is often reduced to binary pairs for easier manipulation in practice. On the other hand, the majority of Greenberg's universals are statistical, because in his statements regarding universals, the expressions "with overwhelmingly

E-mail address: lhtcuc@gmail.com.

[1] According to Lehmann (2005) and Tesnière (1959:32), Schmidt (1926) was the first to use the basic components of the sentence and their interrelationships as a pointer to language typology.

0024-3841/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2009.10.001

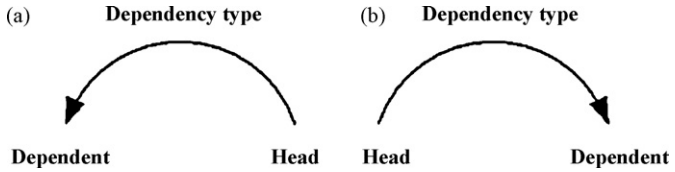

Fig. 1. Three elements of a dependency.

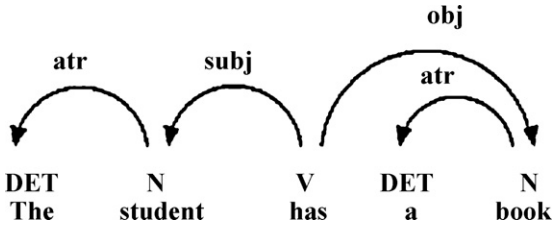

Fig. 2. Syntactic dependency structure of *The student has a book*.

greater than chance frequency" and "dominant" occur with greater than chance probability. Dryer (1998) provides a number of arguments for answering the question of "why statistical universals are better than absolute universals". So, statistical universals are not only useful, but also necessary means for language typology.

If the statistics of binary grammatical pairs and the correlations among them are one of the primary tasks for the study of word-order typology, it is important to choose the best method of building the corpus with the information on grammatical pairs and of extracting such information from the corpus. Compared with previous methods, a corpus-based method can provide more complete and fine-grained typological analysis, while previous methods often focus on basic word order.[2] In this way, the typological conclusion will be based on the language sample used in practice instead of just on some simple sentences collected for the study. A corpus-based method can also ease the task of identifying basic word order, which is often necessary to any linguist working on word-order typology (Song, 2001:49–50).

Following these ideas, this paper proposes a method based on a treebank (namely, a corpus with syntactic annotation), using dependency direction as a typological index, and reporting the results of measuring 20 languages. The paper also presents how to extract typological pairs (SV/VS, VO/OV, AdjN/NAdj) from dependency treebanks and statistically tests the correlation between the ordering of different constituents.

2. Method

In this paper, a dependency approach was employed, according to which the syntactic structure of a sentence consists of nothing but the dependencies between individual words. A detailed discussion on the advantages of the dependency approach is presented in Hudson (2007). The ideas of dependency analysis are found more or less in the traditional grammar of many languages.

The following properties, which are generally accepted by linguists, are considered the core features of a syntactic dependency relation (Mel'čuk, 2003; Ninio, 2006; Hudson, 2007; Liu, 2009a):

1. It is a binary relation between two linguistic units.
2. It is usually asymmetrical and directed, with one of the two units acting as the head and the other as dependent.
3. It is labeled, and the type of the dependency relation is usually indicated using a label on top of the arc linking the two units.

The two units form a dependency pair as shown in Fig. 1.

Fig. 1 shows a dependency relation between Dependent and Head, whose label is dependency type or syntactic function. The directed arc from Head to Dependent demonstrates the asymmetrical relation between the two units. Dependency analysis can be seen as the set of all dependencies found in a sentence. Fig. 2 shows a dependency graph.[3]

[2] Basic word order at the clausal level is found "in stylistically neutral, independent, indicative clauses with full noun phrase (NP) participants, where the subject is defined, agentive and human, the objects is a definite semantic patient, and the verb represents an action, not a state or an event" (Siewierska, 1988:8).

[3] In this analysis, we make the determiner the dependent of the common noun. Some scholars analyze that inversely (Hudson, 2007). Two solutions are acceptable as argued in Hudson (2004).