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

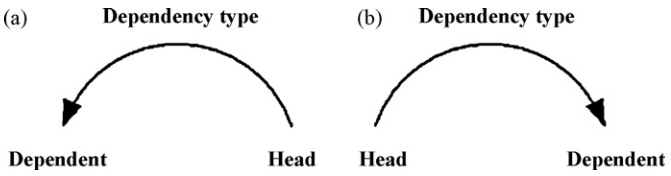

Fig. 1. Three elements of a dependency.

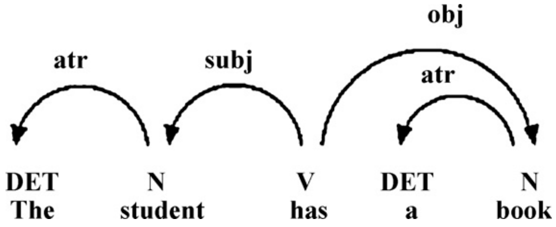

Fig. 2. Syntactic dependency structure of *The student has a book*.

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).

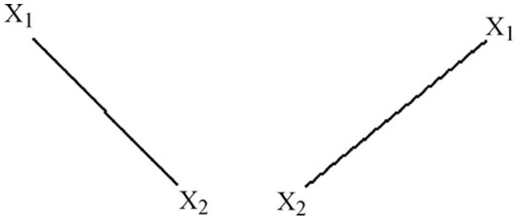

Fig. 3. The linear and structural relations between head and dependent.

Table 1
Annotation of *The student has a book* in the treebank.

Dependent			Head			Dependency type
Order[a] number	Word	POS	Order number	Word	POS	
1	The	det	2	student	n	atr
2	student	n	3	has	v	subj
3	has	v				
4	a	det	5	book	n	atr
5	book	n	3	has	v	obj

[a] The number indicates the linear position of the word in the sentence.

In Fig. 2, all the words in the sentence are connected by grammatical relations. For example, the subject and object depend on the main verb; the determiners depend on the nouns that they modify; and so on.

Fig. 1 also shows that a dependent can precede or follow its head in the linear order of a sentence. We define this phenomenon as dependency direction of a dependency relation. A dependency pair is head-initial if its head precedes the dependent. Inversely, it is head-final if the head follows the dependent.

This idea can be found in Tesnière (1959), which is often considered the founding work of dependency grammar. According to Tesnière (1959:22–23), x1 and x2 are two words in a sentence, and x1 governs x2. If the sequence in the sentence is x1, x2, we have a head-initial order as on the left of Fig. 3. If the sequence in the sentence is x2, x1, we have a head-final order as on the right of Fig. 3.[4]

The differentiation between head-initial and head-final may be important for the classification of languages. According to Tesnière, some languages tend to use more head-final constructions, others use more head-initial ones (1959:32–33).

It is interesting to note the similarity between Tesnière's and Greenberg's concerns about the linear order of two words that form a grammatical relation. The difference between them is also evident: while Greenberg is only concerned with some grammatical relations in a sentence, Tesnière wants to build a full analysis of a sentence based on the binary grammatical relation. Another is the fact that Tesnière also considers the classification of languages based on a head-initial and head-final relation as statistical, because he uses the French word *préférence* (tend to, preference) to describe the tendency of head-initial or head-final word orders in a language.[5]

Following Fig. 1, we can classify dependency relations into head-final (a) and head-initial (b). In terms of current linguistic typology, the former is the OV, and the latter the VO type. In this paper, we also refer to the property of being head-initial or head-final as the dependency direction of a dependency relation.

After having established the distinction between the two dependency relations, our task is to find statistically the distribution of the proportion of the two types of dependency from the corpus, in order to examine whether a language tends to the head-final or head-initial type, and put the results of different languages together in order to determine whether the method can be used for typological purposes.

For finding the preference of a language in terms of head-initial and head-final dependencies, we need to build a sufficiently great corpus, including the sentences with a syntactic structure as in Fig. 2. Such a corpus is also called a treebank (Abeillé, 2003). Treebanks are often used as tools and resources for training and evaluating a syntactic parser in computational linguistics. However, it is obvious that treebanks are also an important tool for basic linguistic concerns (Liu, 2009a,b; Liu et al., 2009).

On the basis of the dependency syntax defined above, we propose the format for a dependency treebank. Table 1 shows the analysis of the sentence in Fig. 2.

In Table 1, each row contains a dependency pair with three elements: dependent, head and dependency type. The linear order of head and dependent are also explicitly marked in all dependencies of a treebank. All rows in a sentence consist of their dependency structure. It can easily be converted into a dependency graph (or tree) as in Fig. 2.

Based on the treebank in Table 1, we can check whether a dependency pair is head-initial or head-final: if the result is a positive number when the order number of the head is deducted from the order number of the dependent, the dependency is

[4] Tesnière uses the term *centrifugal* for *head-initial* and *centripetal* for *head-final*.
[5] Greenberg's universals are stated often in "with overwhelming greater than chance frequency" (Greenberg, 1963).

Table 2
Frequencies and percentage of dependencies of different directions in the Japanese treebank.

	Head-final	Head-initial	Total
Frequencies	97,040	11,937	108,977
Percentage	89%	11%	100%

head-final, otherwise it is head-initial. With this mathematical operation, we can get a frequency distribution of head-final and head-initial dependencies in a sentence or a whole language sample (treebank). For instance, the sentence in Table 1 has three head-final pairs and one head-initial pair.

In the next section, we will present the statistical results of the distribution of head-initial and head-final grammatical pairs for 20 languages.

3. The distribution of head-initial and head-final grammatical pairs in 20 languages

We use the method proposed in section 2 to measure the distribution of dependency direction of 20 languages. These 20 languages are[6]: Chinese (chi), Japanese (jpn), German (ger), Czech (cze), Danish (dan), Swedish (swe), Dutch (dut), Arabic (ara), Turkish (tur), Spanish (spa), Portuguese (por), Bulgarian (bul), Slovenian (slv), Italian (ita), English (eng), Romanian (rum), Basque (eus), Catalan (cat), Greek (ell), Hungarian (hun).

It is evident that the sample is heavily biased toward Indo-European languages (14 of 20 languages), and within Indo-European toward Germanic, Romance, and Slavonic; it is heavily biased toward Europe (18 of 20) and even more so to Eurasia (20 of 20 languages). Our sampling is limited by the available treebanks. We cannot cover many more languages just as in Dryer (1992), because the building of a treebank is a more time-consuming task than of a general typological database. Considering that the aim of this study is to propose a method and check the feasibility of the method to classify the languages, the current sampling seems acceptable.

Most treebanks used in our project are from the training sets of the CoNLL-X Shared Task on Multilingual Dependency Parsing. These treebanks have various annotation schemes, but we use their dependency forms converted by CoNLL-X'06 (Buchholz and Marsi, 2006) and CoNLL-X'07 (Nivre et al., 2007) organizers.[7] Please refer to the Appendix for detailed information on all treebanks used in this study.

Prior to beginning the statistical computing, we have converted all treebanks into the format in Table 1.

We measured the frequencies of head-initial and head-final dependencies for all 20 languages, as shown in Table 2.

Table 2 shows the total number of dependencies as well as the number and percentage of head-initial and head-final dependencies in a treebank. Considering the great difference in size of the 20 treebanks we used, we shall compare them in percentages, which are calculated with the following formulae:

$$\text{percentage of head-final dependency} = \frac{\text{frequencies of the head-final dependency}}{\text{total number of dependencies in the treebank}} \times 100$$

$$\text{percentage of head-initial dependency} = \frac{\text{frequencies of the head-initial dependency}}{\text{total number of dependencies in the treebank}} \times 100$$

Table 3 lists all relevant information of the 20 languages. In the table, *size* is the number of dependencies in this language sample; *msl* is mean sentence length; HF shows the percentage of head-final dependencies in the sample and HI is the percentage of head-initial dependencies; *genre* presents the genre of the sample; *type* shows the native annotation scheme of the treebank, D is dependency structure, C is constituent structure. CF is a mixed structure with constituent and grammatical functions; %n.p. is the percentage of non-projective dependency relations.[8]

We will discuss and analyze the results in the next section.

4. Discussion

For a more lucid view, the distribution of the dependency direction in 20 languages is presented in Fig. 4.

Fig. 4 shows that some languages are more head-initial or head-final than others. There are no pure head-initial or head-final languages in the sample. Each language contains more or less constructions with either order. This finding is similar to that of Dryer (1992), although we used a different method. The figure also points out that most languages favor a balanced

[6] Language codes are following ISO 639-2: Codes for the Representation of Names of Languages. http://www.loc.gov/standards/iso639-2/php/code_list.php.

[7] It is an easy task to convert the format of the CoNLL-X dependency treebank to Table 1. After converting, we can use the method in section 2 to get the needed frequency distribution of dependency directions.

[8] A non-projective dependency relation is a crossing arc in the dependency graph of a sentence. This data in the column is mostly extracted from (Buchholz and Marsi, 2006; Nivre et al., 2007).

Table 3
Statistical overview of dependencies in 20 languages.

	size	msl	HF	HI	genre	type	%n.p.
Arabic (ara)	50,097	35.3	13.3	86.7	News	D	0.4
Bulgarian (bul)	147,071	12.5	31.9	68.1	Mixed	C	0.4
Catalan (cat)	365,530	28.8	38.7	61.3	Mixed	CF	0.1
Chinese (chi)	16,654	24	68.5	31.5	News	D	0.0
Czech (cze)	992,651	14.8	45.5	54.5	News	D	1.9
Danish (dan)	38,120	15.9	19.6	80.4	Mixed	D	1.0
Dutch (dut)	479,677	12.6	48.5	51.5	Mixed	CF	5.4
Greek (ell)	55,953	24.2	50.5	49.5	Mixed	D	1.1
English (eng)	376,563	21.3	51.2	48.8	News	C	0.3
Basque (eus)	47,498	15.8	60.8	39.2	Mixed	D	2.9
German (ger)	564,549	15.4	45.8	54.2	News	CF	2.3
Hungarian (hun)	105,430	21.8	78.1	21.9	News	C	2.9
Italian (ita)	56,822	22.9	35.2	64.8	Mixed	CF	0.5
Japanese (jpn)	108,977	7.9	89	11	Dialog	CF	1.1
Portuguese (por)	168,522	19.6	36.5	63.5	News	CF	1.3
Romanian (rum)	32,108	8.9	33.5	66.5	News	D	0.0
Slovenian (slv)	22,380	15.5	49.8	50.2	Novel	D	1.9
Spanish (spa)	75,571	24	36.4	63.6	Mixed	CF	0.1
Swedish (swe)	160,273	15.5	43.5	56.5	Mixed	D	1.0
Turkish (tur)	38,706	9.3	94.1	5.9	Mixed	D	1.5

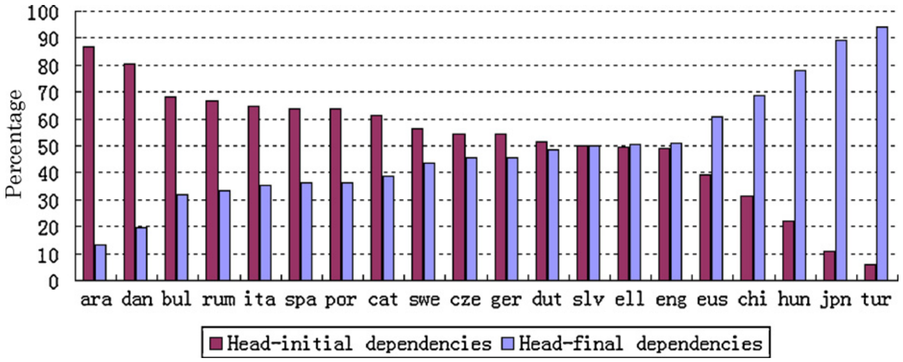

Fig. 4. Distribution of the dependency direction in 20 languages.

distribution of the dependency direction. Being neither strongly head-initial nor strongly head-final, they tend to be a head-medial distribution, if a head has two dependents.

If we observe Fig. 4 from a viewpoint of language typology or classification, at first glance, the dependency direction (head-initial or head-final percentage) can work as a measure for language typology.

Tesnière (1959:33) provides a language classification based on dependency direction. According to his approximate classification,[9] we can arrange the investigated 20 languages in an axis with the two ends being head-first and head-final in Fig. 5.

Comparing Figs. 4 and 5, we find that Arabic, Japanese, Turkish, Chinese and Romance languages have almost perfect consistency between our observation and Tesnière's hypothesis, but there are some differences for Germanic and Slavic languages, in particular, Bulgarian and Danish show the contrary preference.

Bulgarian is not typological with its Slavic kin, but with the Romance languages; this is evidently due to the Balkan linguistic type of Bulgarian, which is not reflected in traditional word-order typology.[10] Danish is wrongly located by its unusual annotation scheme, which is explained in detail in Fig. 6. Based on the treebanks of 20 languages, we empirically prove Tesnière's ideas of language classification, which is a more robust and easily manipulated method. While Tesnière built his classification on a small amount of sentences, the method used in this study has a more modern aspect, which can be applied in automatic means.

Fig. 4 demonstrates a continuum with head-initial and head-final as the ends. Any language can find its position in the continuum. The data also show that a proportion of head-initial and head-final exists and it is different for every language. This implies that languages can be typologized along a continuum and clustered, based on nearness in that continuum, while

[9] Tesnière's classification does not mention all languages in our sample. So, we have to position several languages based on language subgroups (for example, Slavic languages).

[10] We thank an anonymous referee for this observation.

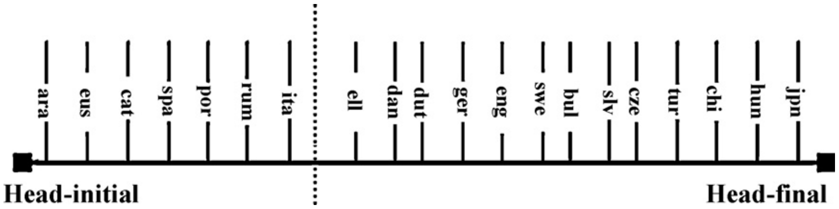

Fig. 5. 20 languages in Tesnière's typological classification system.

Table 4
Statistical overview of several special dependencies in 20 languages. VS, SV, VO, OV, NAdj and AdjN are the percentages (the raw figures are included in parentheses) of the corresponding features in a language; WALS is the dominant word order of the language in Haspelmath et al. (2005). The question mark (?) in the WALS shows that the language has no dominant order[a] in this feature.

	VS	SV	VO	OV	NAdj	AdjN	WALS
Arabic (ara)	61.4 (2153)	38.6 (1351)	91 (5313)	9 (524)	95.9 (3953)	4.1 (167)	VS–VO–NAdj
Bulgarian (bul)	18.5 (3,036)	81.5 (13,417)	90.1 (6224)	9.9 (682)	1.6 (180)	98.4 (11,212)	?–VO–AdjN
Catalan (cat)	18.5 (4584)	81.5 (20,221)	85.5 (19,080)	14.5 (3239)	99.2 (1680)	0.8 (14)	?–VO–NAdj
Chinese (chi)	1.3 (19)	98.7 (1400)	98 (1679)	2 (34)	0.4 (2)	99.6 (461)	SV–VO–AdjN
Czech (cze)	27.4 (34,273)	72.6 (90,841)	72.9 (74,583)	27.1 (27,735)	8.6 (11,521)	91.4 (122,004)	SV–VO–AdjN
Danish (dan)	19.8 (1015)	80.2 (4122)	99.1 (8739)	0.9 (81)	60 (1683)	40 (1124)	SV–VO–AdjN
Dutch (dut)	28.7 (13,258)	71.3 (33,000)	82.5 (71,030)	17.5 (15,085)	7.4 (2024)	92.6 (25,207)	SV–?–AdjN
Greek (ell)	34.7 (1609)	65.3 (3029)	80.5 (3437)	19.5 (834)	8.4 (400)	91.6 (4345)	?–VO–AdjN
English (eng)	3.2 (1116)	96.8 (33,916)	93.5 (28,219)	6.5 (1959)	2.6 (661)	97.4 (24,801)	SV–VO–AdjN
Basque (eus)	20.4 (765)	79.6 (2990)	12.8 (381)	87.2 (2589)	78 (1234)	22 (349)	SV–OV–NAdj
German (ger)	33.2 (17,382)	66.8 (34,938)	36.8 (9447)	63.2 (16,237)	37.1 (15,355)	62.9 (26,016)	SV–?–AdjN
Hungarian (hun)	26.6 (1764)	73.4 (4862)	47.8 (2600)	52.2 (2843)	2.3 (339)	97.7 (14,239)	SV–?–AdjN
Italian (ita)	24.5 (869)	75.5 (2681)	82.3 (2090)	17.7 (451)	60.9 (2374)	39.1 (1523)	?–VO–NAdj
Japanese (jpn)	0	100 (5509)	0	100 (27,553)	0	100 (3820)	SV–OV–AdjN
Portuguese (por)	15.7 (1899)	84.3 (10,190)	85.1 (9447)	14.9 (1656)	70.1 (5858)	29.9 (2495)	SV–VO–NAdj
Romanian (rum)	21.9 (648)	78.1 (2313)	88.3 (1568)	11.7 (208)	66.9 (2905)	33.1 (1439)	SV–VO–NAdj
Slovenian (slv)	38.9 (658)	61.1 (1035)	74.5 (2375)	25.5 (815)	11 (189)	89 (1534)	SV–VO–AdjN
Spanish (spa)	21.5 (1107)	78.5 (4032)	77.3 (3417)	22.7 (1006)	98 (431)	2 (9)	?–VO–NAdj
Swedish (swe)	22.7 (4296)	77.3 (14,589)	94.6 (10,411)	5.4 (596)	0.4 (26)	99.6 (6656)	SV–VO–AdjN
Turkish (tur)	8.1 (284)	91.9 (3208)	4 (255)	96 (6175)	0.3 (11)	99.7 (3514)	SV–OV–AdjN

[a] In fact, here we use *dominant word order* unlike the definition in Croft (2002:60), and closer to the understanding of basic word order in Whaley (1997:100). In other words, it only shows that one of the word order types is more frequent (or dominant) in language use. Dryer (2008a) points out that WALS also uses the *dominant word order* in this meaning, to emphasize that priority is given to the criterion of what is more frequent in language use.

traditional typology uses a classification into a small number of discrete types. Our approach also provides a method to measure the dominant order based on real texts. To determine the dominant word order is a very difficult task, in particular, to distinguish the only order possible or the order that is more frequently used (Dryer, 2008a,b).

Compared with a current study of typology, which is often based on the word order of several special pairs, we extract from 20 treebanks the percentages of the following dependencies: subject–verb, object–verb and adjective–noun. The result is shown in Table 4.

Table 4 shows that the method proposed in this study can be used as a typological means, because the results on dominant word order are very similar with the results in Haspelmath et al. (2005). Therefore, it is reasonable to consider that a dependency treebank can be used as a database for the study of linguistic typology, and as a tool to provide more precise and reliable information to decide which word order is more frequently used in a language.

It is noteworthy that Danish has an opposite result to WALS on the AdjN and NAdj ($p < 0.01$) feature. The problem is caused by a special annotation scheme in the Danish treebank. For instance, an English prepositional phrase is annotated in the Danish treebank as the left side of Fig. 6, while the right side is a more often used annotation (Buch-Kromann, 2006:170).

This very unusual annotation scheme also makes Danish distant from its sisters in the Germanic subgroup and positions it wrongly in the language spectrum (or continuum) as shown in Fig. 4.

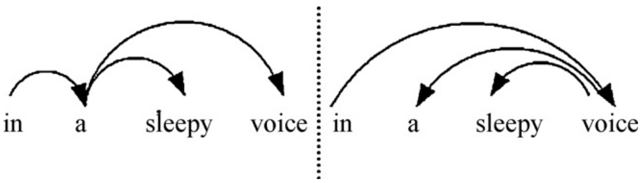

Fig. 6. An English phrase annotated by a Danish treebank scheme.

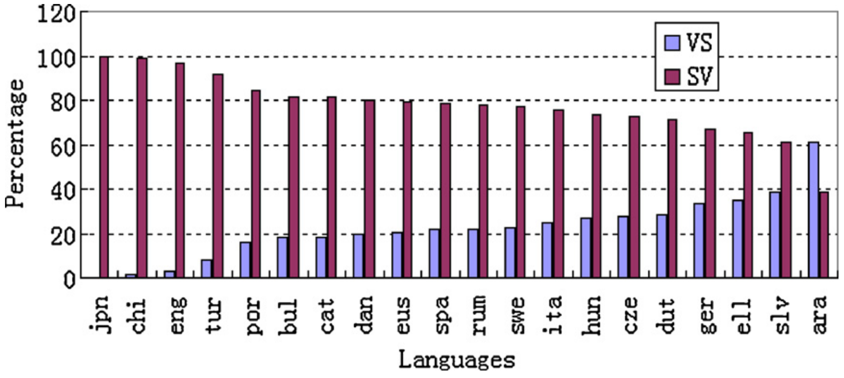

Fig. 7. Distribution of SV and VS in 20 languages.

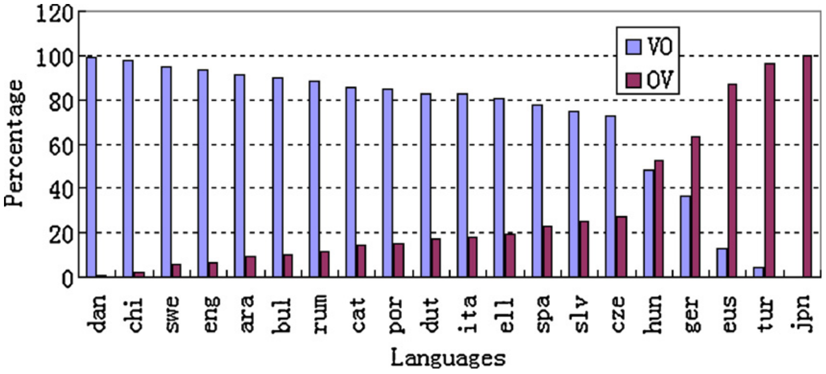

Fig. 8. Distribution of VO and OV in 20 languages.

Now, we turn to the problem of dominant word order. Five languages (Bulgarian, Catalan, Spanish, Italian and Greek) are shown in WALS as having no dominant order for the SV/VS feature. However, the calculations based on Table 4 show that they have a statistically significant ($p < 0.01$) preference for SV.

Fig. 7 shows that all languages other than Arabic favor the SV order. Fig. 8 shows the distribution of VO and OV in 20 languages.

For OV and VO, WALS indicates that three languages (Hungarian, German and Dutch) have no dominant order. It can be seen from Table 4 that Dutch has a preference for VO ($p < 0.01$), and German for OV ($p < 0.01$). However, Hungarian does not have a statistically significant ($p = 0.553$) preference for VO/OV. Therefore, it remains without dominant order on this feature as in WALS. The sources, which make the difference between WALS and our method, are very much worth exploring in future.

While German and Dutch are neighbors in Fig. 7 (SV/VS), they are distant in Fig. 8 (VO/OV), which can even be classified into two contrast types. In the typological literature, there is a tendency to count sentences with S Aux O V_{part} order as SVO. However, in dependency treebanks (German and Dutch), the structure S Aux O V_{part} is analyzed as a dependency graph in Fig. 9.

This influences the counts for VO/OV heavily in two languages, but Fig. 9 cannot explain why Dutch is VO-preferred and German OV-preferred, because Fig. 9 can easily make an OV-preferred language. Two possible factors make Dutch into a VO language: (1) the Dutch treebank includes many sentence analyses uncorrected by human hand, which makes the difference between Dutch and German and (2) the German treebank only uses news genre, while Dutch is based on mixed materials including different genres, which also influences the result (Roland et al., 2007).

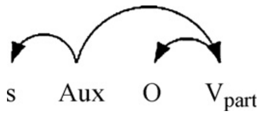

Fig. 9. Dependency graph of a typical sentence in Dutch and German.

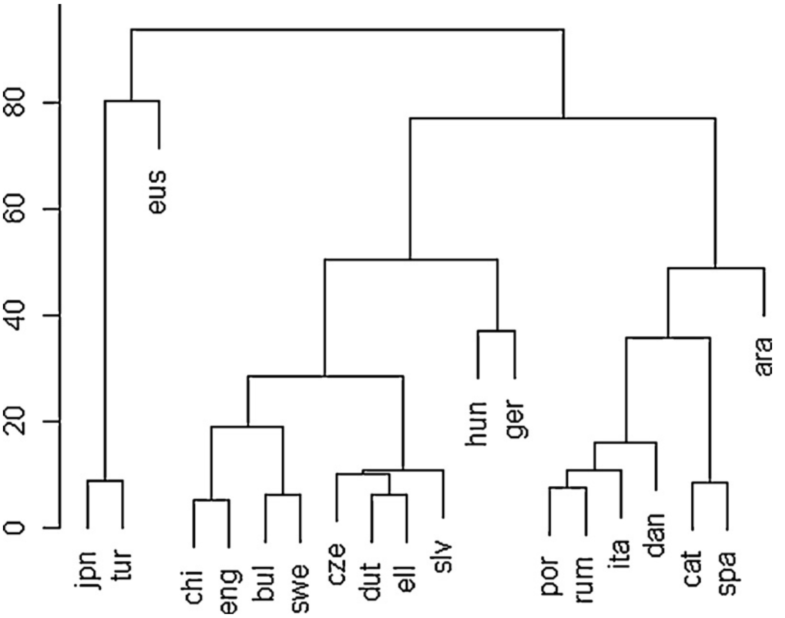

Fig. 10. Clustering of observations for 20 languages.

German and Dutch also raise the following question: how can some special constructions in individual languages be processed, if the method is aimed at a universal measurement for all languages? For instance, for German and Dutch, the most important fact about word order is the verb-second order in main clauses and verb-final order in non-main clauses. Whether and how do we distinguish them in an annotation scheme? To answer this question better, the introduction of two different dependency relations to distinguish the order seems necessary. However, if this is the case, a new question arises: how do we compare the individuality of a language with other languages that do not have these specialties? In the future, perhaps, we should search for some techniques to balance these contrastive relations.

Some differences between two other Germanic languages (e.g. English and German) might be caused by the more fixed word order of English. This property of English is well reflected in its higher percentage of unbalanced VO/OV and SV/VS than German. In this approach, free word order can perhaps be revealed by a more balanced percentage between head-final and head-initial relations, which might be investigated in the future.

The issue of implicational correlations between the ordering of different constituents, whether there are harmonic correlations, is a major issue for word-order typology. It is, therefore, interesting to explore whether we can carry out this correlative study based on dependency treebanks. Following are several correlation tests about the relationship between the order of object and verb and the order of adjective and noun (Dryer, 2008b).

A correlation test between the head-final order in OV/VO dependencies and the head-final order of adjective–noun dependencies in Germanic languages shows that, when Danish is excluded, four Germanic languages have very good harmonic correlations (Pearson correlation coefficient of VO and AdjN = 0.999, p-value <0.001), but when Danish is included, these languages do not correlate in the feature pairs (Pearson coefficient of VO and AdjN = 0.201, p-value = 0.746).

A Pearson correlation test based on a mixed language subgroup was carried out. This subgroup contains the verb–object and adjective–noun languages that are beyond controversy.[11] The result shows that the head-final order in VO dependencies closely correlates with AdjN dependencies (Pearson coefficient = 0.958, p-value < 0.003).

So, it is possible to find or explore the harmonic correlations by the method proposed here.

To compare the approach proposed in this paper with others, some clustering experiments have been conducted. We are applying the “agnes” (**Ag**glomerative **Nes**ting) function[12] in the statistical software R[13] to perform agglomerative hierarchical clustering analyses (Kaufman and Rousseeuw, 1990), using Euclidean distances and the Average method of clustering. Fig. 10 shows a cluster tree of 20 languages based on SV–OV–AdjN features.

The agglomerative coefficient[14] (AC) of the cluster in Fig. 10 is high (0.82), which indicates a good clustering structure. The agglomerative coefficients of other feature combinations are also evaluated by the same method and distance measure. The

[11] The subgroup includes Chinese, Czech, Swedish, Bulgarian, Slovenian, and Greek.

[12] Agnes proceeds by a series of fusions. At first, each observation is a small cluster by itself. Clusters are merged until only one large cluster remains, which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster. If there are several pairs with minimal dissimilarity, the algorithm picks a pair of objects at random.

[13] www.r-project.org (22.04.09).

[14] AC is a dimensionless quantity, varying between 0 and 1. AC close to 1 indicates that a very clear structuring has been found. AC close to 0 indicates that the algorithm has not found a natural structure.

results show that the HF feature has the highest AC (0.92), what remains are HF–VS–VO–NAdj features (0.79), SV–OV features (0.97) and SV feature (0.90).

It is interesting to notice the related languages' position in Figs. 4 and 10. We still use Germanic languages as an example. In Fig. 4, except Danish, which is not in its correct position due to the annotation problems mentioned above, other languages (Swedish, German, Dutch and English) are located very near each other. In Fig. 10, the latter four languages are also distributed in different clusters. We cannot explain why a single factor (percentage of HI or HF) is giving a better result than three parameters (SV/VO/AdjN). Further investigation is needed.

Using a corpus-based method, we have to address issues such as the following: (1) How large a treebank is needed in order to get reliable results? How does the size of the treebank influence the results? (2) What influence has the homogeneity of the corpus? (3) What influence do annotation schemes have on the results?

In order to find the interrelationship among these factors, Liu (2008) investigated five Chinese treebanks of various sizes, annotation schemes and genres. The results show that these factors may influence the result, but they cannot change the conclusions. Because of these factors, annotation scheme and genre are worthy of further discussion here.

Dependency grammar is not a uniform syntactic theory, so we have to consider the issue of whether the annotation scheme is likely to have some influence on the results. The point is that what counts as a dependency relation in one treebank does not automatically carry over to other treebanks. For instance, what is the head and what are the dependents in a coordinating structure? In WH-questions and relative clauses, what is the head? What is the head in a noun phrase? Of all these questions, the one most discussed is, whether in a noun phrase it is the determiner that is dependent on the noun or whether the noun is dependent on the determiner. Both analyses are available in different dependency grammars (Hudson, 2004). In the sample of the present study, the determiner is shown as the head of a noun only in Danish, but in the other languages the noun is chosen as the head of the noun phrase. In fact, the question of how to decide the head in a noun phrase (including the determiner and the noun) has caused one of the few major disagreements in dependency grammar. This uncertainty, even though it is rare in dependency grammar, influences the results of the related treebanks and makes cross-linguistic comparisons unreliable. Thus, we have to inquire whether this influence is likely to affect the conclusion based on the approach proposed here. As the 'type' column in Table 3 shows, in our collection, ten treebanks have been constructed in genuine dependency style, seven were originally in styles of coding function and constituent structure (or constituency), and four were in constituency style. However, all non-native dependency treebanks used in this study have been converted into dependency treebanks. In other words, all treebanks have been brought into alignment with the three core properties of dependency relations mentioned in section 2. It might be preferable to base conclusions only on treebanks using the same annotation schemes. There are four treebanks using the Prague Dependency Treebank annotation scheme: Czech, Arabic, Greek and Slovenian. If the annotation scheme has an important influence on the results, then Czech, Arabic, Greek and Slovenian would have similar characteristics, in contrast with other languages. However, the tables and figures show that these languages are not similar to each other.

Another way to investigate the relationship between annotation schemes and the proposed method is to use several treebanks with different annotation schemes and text genres as a resource for the study of one language. Liu (2008), for example, shows that although the annotation scheme has some influence on the variation in a fraction of head-initial and head-final dependencies, nonetheless, the conclusions from the five treebanks are essentially similar.

Liu (2008) also found that genre has a greater influence than annotation scheme on a fraction of head-initial and head-final dependencies, but the result still falls in the limit, which clearly shows that Chinese is a head-final language. We think that dependency direction is not a sensitive criterion, compared with other grammatical structures in Roland et al. (2007).

Based on these discussions and experiments, it is argued that the annotation scheme and genre of the treebank may influence the results, but that this effect is not strong enough to affect the conclusion seriously. In a study using corpora, a few local properties of a sentence do not suffice to change the global features of a language on the fraction of head-initial and head-final dependencies. However, some extreme cases like Danish also exist, as discussed above. To avoid the interference of these factors on the result, it is reasonable to make some selections related to treebanks before beginning the typological study, based on the method mentioned in the paper. Such a precondition is normal in any scientific work, even when we have a robust tool. A good data resource is always important to get more reliable outcomes.

Another well-known feature that distinguishes languages in dependency treebanks that has attracted a lot of attention from the dependency community is 'projectivity', which was first discussed in Lecerf (1960) and Hays (1964). In a projective dependency graph, no-crossing arcs are allowed. Fig. 11 is a non-projective dependency graph (with crossing arcs).

There are dependency graphs with crossing arcs in some languages. However, no-crossing arcs are often considered as a condition of well-formed dependency graphs for constructing a more efficient parsing algorithm (Nivre, 2006). Fig. 12 presents the percentage of non-projective arcs in our sample.

It is impossible to investigate to what extent non-projective arcs can be used as a typological feature in the current sample, because it is often possible to make a non-projective structure into a projective one by modifying the syntactic annotation scheme. However, some interesting results could be found if we used the same annotation scheme to annotate the corpora of different languages. For instance, Czech, Arabic, Greek and Slovenian are annotated by the same scheme, but Czech and Slovenian have the same percentage, while Arabic and Greek are very different.

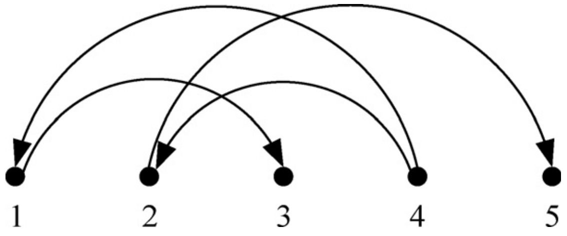

Fig. 11. A dependency graph with crossing arcs.

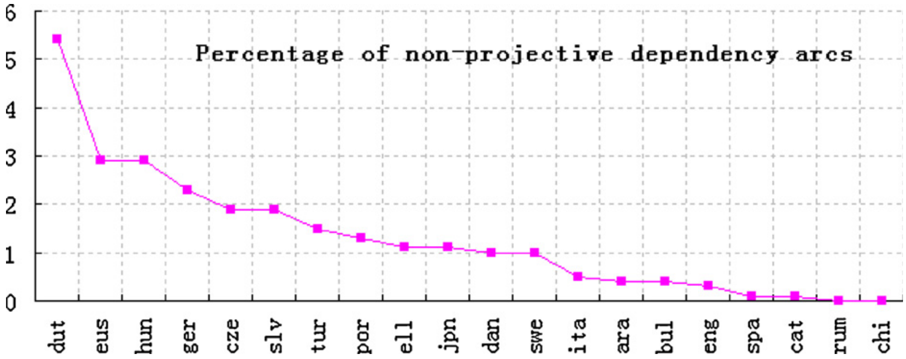

Fig. 12. Non-projective arcs in 20 languages.

5. Conclusions

Compared with the previous quantitative methods in typology (Cysouw, 2005), our method has these advantages and novelties:

(1) it is statistical and corpus-based;
(2) it is robust and non-discrete;
(3) it is more fine-grained;
(4) it is a more overall typological measure for a language;
(5) it can share language resources with computational linguistics.

However, we clearly know that the sample used for the paper is not at all adequate as a typological sample, but it probably is appropriate for the purpose of this study, which demonstrates a method, rather than drawing conclusions about frequencies, distributions, etc. of types. Therefore, there are still many questions open for further research. For example, we do not know why some closely related languages are rather far apart on our cluster cline, and what the smallest size of the treebank is for getting a reliable result.

We shall develop and extend the proposed method in the following ways:

(1) building and using parallel corpora with the same annotation of dependency grammar for different languages;
(2) using the same annotation schemes for the various genres in a language;
(3) trying to find those features, which can be used for the placement of genetically different languages in the same group, i.e. features associated with the dependency direction;
(4) building a larger and better-designed sample of languages and addressing in depth some typological concerns, such as cross-linguistic frequencies of different degrees of head-initial and head-final; continent-to-continent variation; areality and diffusibility; variation within language families; diachronic tendencies; correlations with other parts of grammar, etc.

We believe that these enterprises will be useful for better capturing cross-linguistic similarities and differences, which is one of the primary tasks of modern typology (Bickel, 2007).

Acknowledgments

We thank the anonymous reviewers for valuable suggestions and insightful comments, Richard Hudson and Michael Cysouw for insightful comments for earlier versions, Rudolf Salzlechner and Li Minglin for improving the English, the organizers and all treebank providers of CoNLL-X 2006 and 2007 dependency parsing for the treebanks. This work is partly supported by the National Social Science Foundation of China (Grant No. 09BYY024) and Communication University of China as one of "211" key projects.

Appendix A

Prague Dependency Treebank (PDT, Czech); Prague Arabic Dependency Treebank (PADT); Slovene Dependency Treebank (SDT); Danish Dependency Treebank (DDT); Swedish Talbanken05; Turkish Metu-Sabanči treebank; German TIGER treebank; Japanese Verbmobil treebank; The Floresta sintá(c)tica (Portuguese); Dutch Alpino treebank; Spanish Cast3LB; Bulgarian BulTreeBank; Romanian dependency Treebank[15]; English Penn Treebank, CuC Chinese dependency treebank; Italian Syntactic-Semantic Treebank (ISST); Basque Treebank; CESS-Cat Catalan treebank[16]; Szeged Treebank (SzTB)[17]; Greek Dependency Treebank (GDT). These treebanks are described in the following documents.

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., Diaz de Ilarraza, A., Garmendia, A., Oronoz, M., 2003. Construction of a Basque dependency treebank. TLT 2003. Second Workshop on Treebanks and Linguistic Theories, Vaxjo, Sweden, November 14–15.

Afonso, S., Bick, E., Haber, R., Santos, D., 2002. "Floresta sinta(c)tica": a treebank for Portuguese. In: Proc. of LREC-2002, pp. 1698–1703.

Atalay, N.B., Oflazer, K., Say, B., 2003. The annotation process in the Turkish treebank. In: Proc. of LINC-2003.

Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G., 2002. The TIGER treebank. In: Proc. of TLT-2002.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A., 2005. The Szeged treebank. In: Matousek, V., et al. (Eds.), Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005, LNAI 3658. Springer Verlag, pp. 123–131.

Civit Torruella, M., Marťi Antoňin, Ma.A., 2002. Design principles for a Spanish treebank. In: Proc. of TLT-2002.

Dzeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Zabokrtsky, Z., Zele, A., 2006. Towards a Slovene dependency treebank. In: Proc. of LREC-2006.

Hajic, J., Smrz, O., Zemanek, P., Snaidauf, J., Beska, E., 2004. Prague Arabic dependency treebank: development in data and tools. In: Proc. of NEMLAR-2004, pp. 110–117.

Kawata, Y., Bartels, J., 2000. Stylebook for the Japanese treebank in VERBMOBIL. Verbmobil-Report 240, Seminar fur Sprachwissenschaft, Universitat Tubingen.

Liu, H., 2007. Building and using a Chinese dependency treebank. grkg/Humankybernetik 48(1), 3–14.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., Delmonte, R., 2003. In: Abeille, A. (Eds.), Building the Italian syntactic-semantic treebank. pp. 189–210.

Nilsson, J., Hall, J., Nivre, J., 2005. MAMBA meets TIGER: reconstructing a Swedish treebank from antiquity. In: Proc. of the NODALIDA Special Session on Treebanks.

Oflazer, K., Say, B., Zeynep Hakkani-Tur, D., Tur, G., 2003. Building a Turkish treebank. In: Abeille (2003), chapter 15.

Prokopidis, P., Desypri, E., Koutsombogera, M., Papageorgiou, H., Piperidis, S., 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In: Civit, M., Kübler, S., Martí, Ma.A. (Eds.), Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005), Universitat de Barcelona, Barcelona, Spain, December 2005, pp. 149–160.

Simov, K., Osenova, P., 2003. Practical annotation scheme for an HPSG treebank of Bulgarian. In: Proc. of LINC-2003, pp. 17–24.

van der Beek, L., Bouma, G., Malouf, R., van Noord, G., 2002. The Alpino dependency treebank. In: Computational Linguistics in the Netherlands (CLIN).

References

Abeillé, A. (Ed.), 2003. Treebanks: Building and Using Parsed Corpora. Kluwer Academic Publishers, Dordrecht.

Bickel, B., 2007. Typology in the 21st century: major current developments. Linguistic Typology 11 (1), 239–251.

Buchholz, S., Marsi, E., 2006. CoNLL-X shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), New York City, June 2006, pp. 149–164.

Buch-Kromann, M., 2006. Discontinuous grammar. A dependency-based model of human parsing and language acquisition. Dr.ling.merc. dissertation, Copenhagen Business School, July 2006.

Croft, W., 2002. Typology and Universals, 2nd ed. Cambridge University Press, Cambridge.

Cysouw, M., 2005. Quantitative methods in typology. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Hrsg.), Quantitative Linguistik. Ein internationales Handbuch (Quantitative Linguistics. An International Handbook). de Gruyter, Berlin, New York, pp. 554–578.

Dryer, M.S., 1992. The Greenbergian word order correlations. Language 68, 81–138.

Dryer, M.S., 1997. On the 6-way word order typology. Studies in Language 21, 69–103.

Dryer, M.S., 1998. Why statistical universals are better than absolute universals. Chicago Linguistic Society: The Panels 33, 123–145.

Dryer, M.S., 2008a. Order of degree word and adjective. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich, chapter 91. Available online at http://wals.info/feature/91 (accessed 16.04.09).

[15] http://phobos.cs.unibuc.ro/roric/texts/indexen.html.

[16] http://www.lsi.upc.edu/~mbertran/cess-ece.

[17] http://www.inf.u-szeged.hu/hlt.

Dryer, M.S., 2008b. Relationship between the order of object and verb and the order of adjective and noun. In: Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich, chapter 91. Available online at http://wals.info/feature/91 (accessed 16.04.09).

Greenberg, J.H., 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. (Ed.), Universals of Language. MIT Press, Cambridge, MA, pp. 58–90.

Haspelmath, M., Dryer, M., Gil, D., Comrie, B. (Eds.), 2005. The World Atlas of Language Structures. Oxford University Press, Oxford.

Hays, D.G., 1964. Dependency theory: a formalism and some observations. Language 40, 511–525.

Hudson, R., 2004. Are determiners heads? Functions of Language 11, 7–43.

Hudson, R., 2007. Language Networks: The New Word Grammar. Oxford University Press, Oxford.

Kaufman, L, Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Lecerf, Y., 1960. Programme des conflits-modèle desconflits. Rapport CETIS, No. 4, Euratom, pp. 1–24.

Lehmann, W.P., 2005. A note on the online publication of Proto-Indo-European Syntax. http://www.utexas.edu/cola/centers/lrc/books/piesN.html (05.05.07).

Liu, H., 2008. A quantitative study of Chinese structures based on dependency treebanks. Changjiang Xueshu 3, 120–128 in Chinese.

Liu, H., 2009a. Dependency Grammar: from theory to practice. Science Press, Beijing.

Liu, H., 2009b. Probability distribution of dependencies based on chinese dependency treebank. Journal of Quantitative Linguistics 16 (3), 256–273.

Liu, H., Hudson, R., Feng, Zh., 2009. Using a Chinese treebank to measure dependency distance. Corpus Linguistics and Linguistic Theory 5 (2), 161–174.

Mel'čuk, I., 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. In: Agel, V., Eichinnger, L., Eroms, H.-W., Hellwig, P., Herringer, H. J., Lobin, H. (Eds.), Dependency and Valency. An International Handbook of Contemporary Research, vol. 1. W. de Gruyter, Berlin/New York, pp. 188–229.

Ninio, A., 2006. Language and the Learning Curve: A New Theory of Syntactic Development. Oxford University Press, Oxford.

Nivre, J., 2006. Inductive Dependency Parsing. Springer Verlag, Dordrecht.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D., 2007. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. pp. 915–932.

Roland, D., Dick, F., Elman, J., 2007. Frequency of basic English grammatical structures: a corpus analysis. Journal of Memory and Language 57 (3), 348–379.

Schmidt, W., 1926. Die Sprachfamilien und Sprachenkreise der Erde. Carl Winter Universitätsverlag, Heidelberg.

Siewierska, A., 1988. Word order rules. Croom Helm, London.

Song, J., 2001. Linguistic Typology: Morphology and Syntax. Pearson Education, Harlow/London.

Tesnière, L., 1959. Eléments de la syntaxe structurale. Klincksieck, Paris.

Whaley, L.J., 1997. Introduction to Typology: The Unity and Diversity of Language. Sage Publications, New York.