

Dependency Relations and Dependency Distance - a statistical view based on Treebank

Haitao Liu

Institute of Applied Linguistics
Communication University of China
CN-100024, Beijing, China
lhucuc@gmail.com

Abstract

The dependency relation is the most essential ingredient in a dependency-based theory of syntax. This paper presents some statistical findings on the dependency relation extracted from a Chinese dependency treebank. A sentence in the proposed treebank can easily be converted into a SSyntS graph in Meaning-Text Theory. The statistics on the dependency relation show that modifiers make up 55% of all dependencies and actants have a lower proportion of 45%. The paper demonstrates it is possible to extract from the treebank active and passive valence information of a word (or word class). The paper gives a formula to calculate the mean dependency distance (MDD) for a specific type of dependency relation in a language and obtains MDD of all dependency types in Chinese. These figures show that some dependencies tend to be much farther apart than others, and demonstrate that dependency distance tends to minimization and different dependency types have varying preference on the direction of dependency.

Keywords

dependency syntax, dependency relation, dependency distance, dependency direction, Chinese treebank, quantitative analysis

1 Introduction

Sentence analysis based on the dependency relations has a longer history than the method based on phrase structure. The ideas of dependency analysis are found more or less in the traditional grammar of many languages. Linguists still have different understandings for what the dependency relation is, but the following properties, which are generally accepted by linguists, are considered the core features of a syntactic dependency relation (Mel'čuk, 2003; Nivre, 2006; Hudson, 2007):

1. It is a binary relation between two linguistic units.

2. It is usually asymmetrical, with one of the two units acting as the governor and the other as dependent.
3. It is labeled, so the dependency relations should be distinguished and explicitly labeled in the arc linking the two units.

Dependency is a core operation for any dependency-based grammar (Ninio, 2006: 8). For the Meaning-Text Model, our study relates to syntagms or Synt-D in the Surface-Syntactic Component and can provide for the writer of local rules (syntagms) the quantitative data regarding the basic dependency structure of a language. Such study is necessary because of “the syntagms being the backbone of any syntactic description”(Mel’čuk & Pertsov, 1987: 182-283) and “Synt-Ds are building blocks of a SyntS”(Mel’čuk, 2003: 198).

Treebank is a corpus with syntactic annotation. It is often used as a tool and a resource for training and evaluating a syntactic parser in computational linguistics (Abeillé, 2003). However, treebanks are not only useful to computational linguists, they are also an important tool for syntacticians to extract information on a language’s structure.

This paper will present some statistical methods and results on dependency relation extracted from a Chinese dependency treebank which is built based on the three properties of the dependency relation.

Section 2 introduces the format of the treebank used, and other related information about it. Section 3 shows some statistical results of dependency relations based on the treebank. Section 4 calculates the dependency distance of the dependency types in the treebank. Section 5 presents concluding remarks and directions of further work.

2 Chinese dependency treebank

The dependency syntax of a language contains two parts: the tagset of word classes, and the tagset of dependency types. Based on the national standard of China “POS tagset for Chinese information processing” and popularly used “Grammar system for middle-school teaching”, we propose a set of word class with 13 main types. The dependency tagset contains 20 SSynt-actants (complements) and 14 SSynt-modifiers (adjuncts). The repertoire of Chinese actants is a little larger than that of other languages (Maxwell & Schubert, 1989) because Chinese has to use functional words for expressing grammatical functions which in other languages are often morphologically realized.¹

On the basis of the dependency syntax defined above, in Table 1 we propose the format for a Chinese dependency treebank². Table 1 shows the analysis of an example in terms of this dependency syntax, with each word token distinguished by a number showing the linear order of the word in the sentence.

¹ For more details on the Chinese dependency syntax, see (Liu & Huang, 2006). (Liu, 2007a) presents how to use the same treebank for the application of computational linguistics.

² In the Table 1, *r* is a pronoun, *v* is a verb, *m* is a numeral, *q* is a classifier, *n* is a noun, *bjd* is an end punctuation of the sentence. The meaning of the abbreviations of dependency types are given in Table 3.

Order number of Sentence	Dependent			Governor			Dependency type
	Order number	Character	POS	Order number	Character	POS	
S1	1	这	r	2	是	v	subj
S1	2	是	v	6	。	bjd	s
S1	3	一	m	4	个	q	qc
S1	4	个	q	5	例子	n	atr
S1	5	例子	n	2	是	v	obj
S1	6	。	bjd				

Table 1: Annotation of a sample sentence in the treebank

This format includes all three mentioned elements of the dependency relation, and can easily be converted into an SSyntS graph (or tree) as in Figure 1.

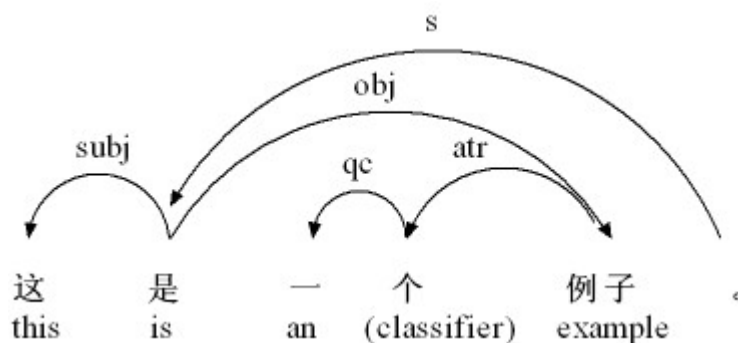


Figure 1: The dependency analysis of a sentence as a graph

Using the proposed Chinese dependency syntax and treebank’s format, we have built a Chinese treebank whose annotation material is the news (xinwen lianbo) of China Central Television, a genre which is intended to be spoken but whose style is similar to the written language. The final treebank includes 711 sentences and 20,034 word tokens; the mean sentence length is 28 words.

3 Statistics of Dependencies in a Chinese Treebank

Excluding punctuations, there are 17,809 dependencies in the corpus, covering 32 dependency types of the syntax. The frequency distribution of the 32 types is shown in Figure 2.

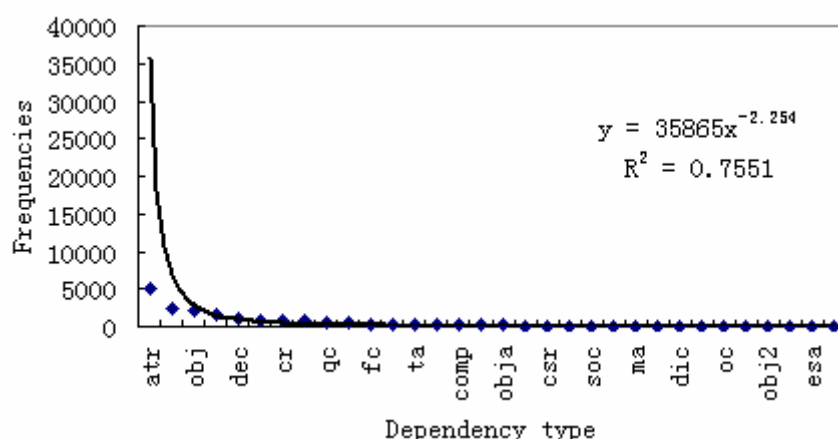


Figure 2: Frequency distribution of 32 dependency types in the treebank. The equation of the regression and the determination coefficient R^2 are indicated in the right-upper region.

It is noteworthy that modifiers make up 55% of all dependency relations and actants have a smaller proportion of 45%. Figure 2 also shows that it is impossible to build a working grammar based on a model which only includes actants. In Figure 2, the fitting with Zipf-like power-law curve is not as good as in other examples in linguistics (Baayen, 2001; Ninio, 2006). Perhaps, too few dependency types make the plot deviate from the perfect result, or, the distribution of dependency relations simply does not follow the known law. We also apply the method to a greater Chinese dependency treebank with 200K dependencies. The results are similar to the results from this one.

	Dependent			Governor		
	noun	pronoun	other	verb	adjective	other
<i>subject</i>	78.37%	8.81%	12.82%	92.53%	3.52%	3.95%
<i>object</i>	70.81%	0.58%	28.61%	97.02%	0	2.98%

Table 2: The distribution of dependency types *subject* and *object*

Table 2 reveals that in the news genre, the noun is the principal word class as *subject*, the pronoun is in the second position, but the function of pronoun as *object* is extremely rare. Our conclusions based on the Chinese treebank are similar to (Biber et al., 2000: 236) on English.

Based on treebank, it is relatively easy to extract all governable dependency types of a word class (*active valence* in Mel'čuk & Pertsov, 1987: 80). Figure 3 shows such feature of a verb in Chinese. In these dependency types, 24.77% is *adva*, 19.1% *obj*, 15.09% *subj*, 11.09% *punct*, 7.93% *cr*, 4.68% *va*, 2.4% *ta*, and 1.6% *comp* etc.

The treebank also can provide the information by which word classes a particular word class is governed (*passive valence*). In summary, the treebank gives us the means to extract syntactic valence information of a word (class). Such information is useful and necessary to build *government patterns* of a word (or word class) in a MTT-style lexicon.

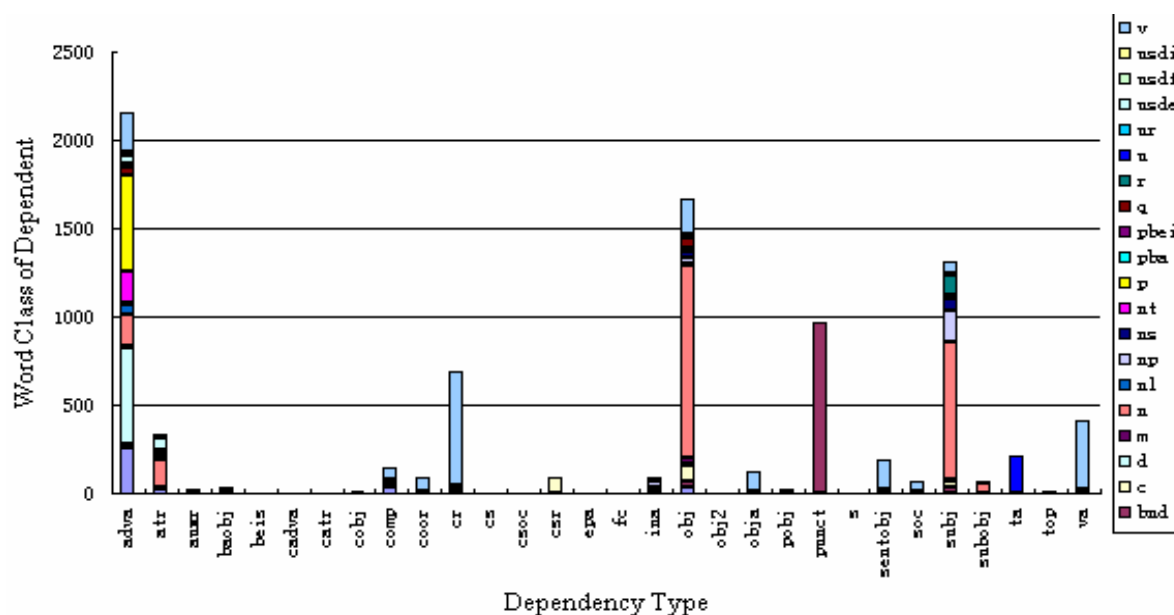


Figure 3: Dependency types governed by a verb

The figures from the treebank also point to possible directions of further research. For example, it is relatively obvious that nouns can play the role of *subj*, *obj*, *pobj*, *epa* and *dec*, but the data also show that 31.1% of the nouns are governed by other word classes, chiefly nouns through the dependency type *atr*. It is an open question to be explored whether this is a result influenced by stylistic differences or a universal phenomenon.

4 Dependency Distance

Dependency distance is the linear distance between governor and dependent. The term ‘dependency distance’ was introduced in (Hudson, 1995:16) and is further discussed in (Hudson, 2003; 2007). The study of dependency distance (DD) is useful for: (1) Predicting syntactic difficulty (Gibson, 1998); (2) Recognizing the mechanisms of children's language learning (Ninio, 2006); (3) Designing better parsing algorithms for natural language processing (Collins, 1996; Buch-Kromann, 2006).

(Liu, 2006) proposes a measuring method of dependency distance based on the treebank and works out that the mean dependency distance of all dependencies in the Chinese treebank is 2.81. (Liu, 2007b) shows that probability distribution of dependency distance in a text can be well captured by the right truncated Zeta distribution. In this paper, we principally investigate dependency distance of dependency relations (types), for instance, the dependency distance of *subject* or *adverbial* from their governor.

Formally, let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_a and W_b , if W_a is governor and W_b is dependent, then the dependency distance (DD) between them can be defined as the difference $a-b$. When a is greater than b , the DD is a positive number, which means that the governor occurs after the dependent; when a is smaller than b , the DD is a negative number and the governor precedes the dependent. It is noteworthy that our method not only can be used to calculate the dependency distance, it is

also useful to determine the direction of a dependency relation. This is the reason why we, unlike others, always count the distance between two words as at least 1. The formula (1) allow us to calculate the mean dependency distance (MDD) for a specific type of dependency relation in a language sample (treebank):

$$(1) \quad MDD(\text{dependency type}) = \frac{1}{n} \sum_{i=1}^n DD_i$$

Here n is the number of examples of that relation in the sample. DD_i is the dependency distance of the i -th syntactic link in the set of that dependency type. If the result is a positive number, it means that the dependency type is a governor-final dependency type. If a negative number, the dependency type is governor-initial. The result is also easily to convert to the explanation of dependency distance defined by Hudson as “the distance between words and their parents, measured in terms of intervening words.”(Hudson, 1995:16) For instance, if we get a value -2.5, which means that the dependency type is governor-initial and there are approximately 1.5 words between the governor and the dependent. (Kahane, 2001: 22) proposes using *local order constraints* in the syntactic-to-morphology correspondence under MTT. He gives weights for the dependencies and uses the signs (-) or (+) to indicate if the dependent is before or after the governor, and the absolute value of the weight to indicate the relative distance between the dependent and the governor. Obviously, measuring dependency distance based on the treebank can provide more precise data for the *local order constraints* and the linearization of the SSyntS. It is also useful in text generation to determine the word order and in parsing for disambiguating. (Holan et al., 1998) presents two measures of non-projectivity in a dependency-based formal grammar for formulating constraints on the degrees of word-order freedom in a language. Our method can also provide some empirical data for such theoretical study. We used formula (1) to calculate MDD of all dependency types in the treebank. Some results are found in Table 3. Table 3 reveals some interesting phenomena. These figures need more research but they show very clearly that some dependencies tend to be much farther away than others. In the following, we will try to speculate on the possible causes that make for the differences.

1. MDD of *subject* is 2.8, and of *object* is -3.94. The difference reflects the unbalanced structure of a Chinese sentence. Perhaps it is explainable from the viewpoint of communicative structure. When the subject appears, it has to be soon followed by its governor, in order to get the basic meaning of the sentence. If the governor of object was presented, the basic meaning of the sentence has been built. In that time, the object can appear a little late.
2. MDD of *adverbial* and *attribute* are compatible with our intuition on Chinese structure. They often appear before their governors, and *attribute* is closer than *adverbial* to the governor.
3. MDD of *classifier complement*, *complements* and *aspect adjunct* prove the adjacent character of these relations with their governors.
4. MDD of *sentential object*, *parenthesis*, *clausal relation* show that they are not central elements in a sentence.

Dependency Relations and Dependency Distance

5. The sign and value of MDD of *main governor* demonstrates that Chinese tends to use, in a sentence including several clauses, the first verb as the main governor of the sentence.

Dependency type	Label	Dependencies	Percentage	MDD
Main governor	s	753	4.23	19.54
Parenthesis	ina	95	0.53	7.8
Correlative adjunct	csr	100	0.56	4.36
Adverbial	adva	2412	13.54	3.46
Subject	subj	1602	9	2.8
Complement of usde	dec	1182	6.64	2.21
Epithet	epa	278	1.56	1.67
Attributer	atr	5016	28.17	1.62
Postpositional Complement	fc	284	1.59	1.45
Complement of classifier	qc	479	2.69	1.13
Complement	comp	192	1.08	-1.01
Aspect adjunct	ta	217	1.22	-1.08
Coordinating adjunct	coor	209	1.17	-1.33
Verb adjunct	va	470	2.64	-2.98
Prepositional Object	pobj	863	4.85	-3.67
Object	obj	2067	11.61	-3.94
Sentential object	sentobj	187	1.05	-7.96
Clause relation	cr	803	4.51	-10

Table 3: Dependency distances and distribution or some dependency types in a Chinese treebank. “Dependency type” and “Label” for the name and tag of the related dependency type; “Dependencies” for the frequencies of the dependency and “Percentage” for the percentage of the dependency in all dependencies in the treebank; “MDD” for mean dependency distance of all dependencies of the dependency type.

The distribution of dependency distance of a dependency type also can be seen more clearly in a time series plot of all dependencies with the same dependency label. Figure 4 shows the distribution of dependency distances of a given dependency type.

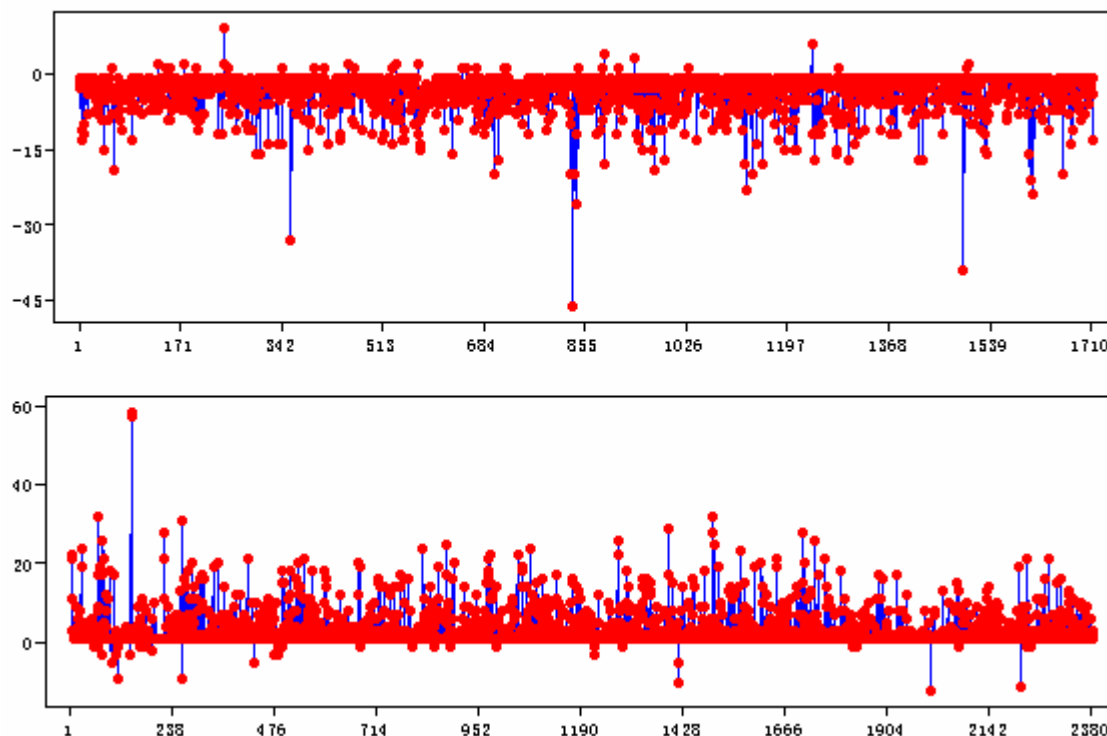


Figure 4: Time series plot of the dependency types *object*(above) and *adverbial*(below).

Y-axis is dependency distance with direction and X-axis is the index number of the dependency in all dependencies of the type.

The distribution in Figure 4 illustrates that the dependency distance of a dependency type is closely related to its statistical aspect. (Ferrer i Cancho, 2004), based on a Romanian dependency treebank, makes hypotheses that the average distance of a sentence is minimized. The density around zero in Figure 4 shows that dependency distance tends to minimization even in a specific type of dependency relation in a treebank. Figure 4 demonstrates that different dependency types have a strong preference for a given dependency direction. In other words, if we investigate the linear relation between dependent and governor, not only the distance should be considered; the direction (or precedence relation) is also another important factor. (Courtin & Genthial, 1998) also emphasize the importance of the relative position between the dependent and its governor from the viewpoint of parsing.

5 Conclusions

Treebanks are not only useful to computational linguists, they are also an important tool for syntacticians. In this paper we propose a format of treebank that can easily be converted into a SyntS graph in Meaning-Text Theory. The paper presented some statistical methods and results extracted from a Chinese dependency treebank with the three properties of dependency relations. The statistics on the distribution of dependency relations in a large corpus show that modifiers make up 55% of all dependency relations and actants have a smaller proportion of

45%. The data is useful to show that it is impossible to build a working grammar completely based on actants. We also showed that it is possible to extract from treebank valency information for a word or for a word class. The paper gives a formula to calculate the mean dependency distance (MDD) for a specific type of dependency relation in a language sample and how to obtain MDD of all dependency types in the treebank. These figures not only show very clearly that some dependencies tend to be much more distant than others, but also demonstrate that dependency distance tends to minimization, and that different dependency types have a preference for a specific direction of dependency. We have also applied the proposed method to a greater Chinese dependency treebank with 200K dependencies, with conclusions similar to the present ones.

Considering the importance of the dependency relation for any linguistic theory based on the dependency principle, the study contributes to a clearer understanding of the essentials of dependency. It is also noteworthy that the figures presented in this paper may be influenced by the size and genre of the treebank. For finding the interrelations among these factors, we are working on five Chinese dependency treebanks with various genres, sizes and annotation schemes. We are also doing a cross-linguistic comparative study on dependency distance and dependency direction of 20 languages based on the method proposed in this paper.

Acknowledgements

I am grateful to Anat Ninio for providing detailed comments and improving my English, Richard Hudson and Feng Zhiwei for insightful discussion, Bernd Bohnet for closely linking the paper to MTT, two anonymous reviewers for helpful comments, and Zhao Yiyi for annotating treebank. Work described in this paper was partly supported by The State Administration of Radio, Film & TV of China (the research project BW0357).

Bibliography

- Abeillé, A. (ed). 2003. *Treebank: Building and using Parsed Corpora*. Dordrecht: Kluwer.
- Baayen, R. H. 2001. *Word Frequency Distribution*. Dordrecht: Kluwer.
- Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 2000. *Longman Grammar of Spoken and Written English*. Beijing: FLTRP.
- Buch-Kromann, M. 2006. *Discontinuous Grammar. A dependency-based model of human parsing and language acquisition*. Dr.ling.merc. dissertation, Copenhagen Business School.
- Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the Association for Computational Linguistics*, 184-191.
- Courtin, J. & D. Genthial. 1998. Parsing with Dependency Relations and Robust Parsing. In Kahane S. & Polguère A. (eds), *Workshop on dependency-based grammars, COLING-ACL'98*. Montréal. 95-101.
- Ferrer i Cancho, R. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70, 056135.

- Gibson, E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1-76.
- Holan, T., V. Kubon, K. Oliva & M. Plátek. 1998. Two Useful Measures of Word Order Complexity. In Kahane S. & Polguère A. (eds), *Workshop on dependency-based grammars, COLING-ACL'98*. Montréal. 21-28.
- Hudson, R. A. 1995. Measuring Syntactic Difficulty. Unpublished paper.
<http://www.phon.ucl.ac.uk/home/dick/difficulty.htm> (2007-3-31)
- Hudson, R. A. 2003. The psychological reality of syntactic dependency relations. In *Proceedings of First International Conference on Meaning-Text Theory*. Paris. 181-192.
- Hudson, R. A. 2007. *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
- Kahane, S. 2001. A fully lexicalized grammar for French based on Meaning-Text theory. In *Computational Linguistics, Proc. CICLing 2001*, Mexico, 18-31. Springer Verlag.
- Liu, H. 2006. Syntactic parsing based on dependency relations. *Grkg/Humankybernetik*, 47(3):124-135.
- Liu, H. 2007a. Building and using a Chinese dependency treebank. *Grkg/Humankybernetik*, 48(1): 3-14.
- Liu, H. 2007b. Probability distribution of dependency distance. *Glottometrics* 15: 1-13.
- Liu, H. & W. Huang. 2006. A Chinese Dependency Syntax for Treebanking. In *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation*. Wuhan. 126-133. Beijing: Tsinghua University Press.
- Maxwell, D. & K. Schubert, (ed). 1989. *Metataxis in practice: dependency syntax for multilingual machine translation*. Dordrecht: Foris.
- Mel'čuk, I. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, L. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1. Berlin - New York: W. de Gruyter, 188-229.
- Mel'čuk, I. & N. Pertsov, 1987. *Surface Syntax of English. A Formal Model within the Meaning-Text Framework*. Benjamins, Amsterdam.
- Ninio, A. 2006. *Language and the Learning Curve: A new theory of syntactic development*. Oxford: Oxford University Press.
- Nivre, J. 2006. *Inductive Dependency Parsing*. Dordrecht: Springer.