Dependency Linguistics

Recent advances
in linguistic theory using
dependency structures

Edited by
Kim Gerdes
Eva Hajičová
Leo Wanner

Dependency Linguistics

Recent advances in linguistic theory
using dependency structures

Edited by

Kim Gerdes
University Sorbonne Nouvelle Paris 3

Eva Hajičová
Charles University Prague

Leo Wanner
ICREA and Pompeu Fabra University Barcelona

∞ ™ The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI z39.48-1984.

Dependency Linguistics

Linguistik Aktuell/Linguistics Today (LA)

Linguistik Aktuell/Linguistics Today (LA) provides a platform for original monograph studies into synchronic and diachronic linguistics. Studies in LA confront empirical and theoretical problems as these are currently discussed in syntax, semantics, morphology, phonology, and systematic pragmatics with the aim to establish robust empirical generalizations within a universalistic perspective.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/la

Volume 215

Dependency Linguistics. Recent advances in linguistic theory using dependency structures
Edited by Kim Gerdes, Eva Hajičová and Leo Wanner

Authors

Margarita Alonso Ramos
Universidade da Coruña

Markus Dickinson
Indiana University, Bloomington

Eva Duran Eppler
University of Roehampton, London

Pavlína Jínová
Charles University in Prague

Dina El Kassas
Minya University

Thomas Groß
Aichi Uninversity, Toyohashi, Japan

Iørn Korzen
Copenhagen Business School

Nicolas Mazziotta
Universität Stuttgart

Igor Mel'čuk
Université de Montréal

Jiří Mírovský
Charles University in Prague

Henrik Høeg Müller
Copenhagen Business School

Kristiina Muhonen
University of Helsinki

Timothy Osborne
Kirkland, WA, USA

Jarmila Panevová
Charles University in Prague

Lucie Poláková
Charles University in Prague

Andreas Pankau
Goethe Universität Frankfurt am Main & Universiteit Utrecht

Tanja Purtonen
University of Helsinki

Marwa Ragheb
Indiana University, Bloomington

Kateřina Rysová
Charles University in Prague

Magda Ševčíková
Charles University in Prague

Orsolya Vincze
Universidade da Coruña

Foreword

Kim Gerdes, Eva Hajičová & Leo Wanner
Sorbonne Nouvelle / Charles University / ICREA and Pompeu Fabra University

1. Is dependency a linguistic domain?

Are you a dependency linguist? Is Dependency Linguistics a field of linguistics at all? It certainly is a field that is different from syntax or language acquisition in that it does not have a delimited part of language that it tries to shed light on. It rather resembles Computational Linguistics that is defined by its primary tool, computers, because Dependency Linguistics is defined by the primary linguistic representation it uses: dependency structures, i.e. hierarchical relations primarily between the basic units of language (words, morphemes, semantemes, illocutionary units, etc.). The links are often labeled and categorized, and the units and links together commonly form a graph, usually a directed acyclic graph, sometimes even a tree or a chain. You are thus a dependency linguist if you assign a prominent role to these kinds of structures in your analyses – be it in syntax, semantics, discourse analysis, or any other field of language sciences.

The choice of dependency graphs as the primary representation opposes Dependency Linguistics directly to constituent analyses, where the main goal is to define groupings, commonly continuous chunks of text that result in constituent or even phrase structure trees, if the groupings can be hierarchically organized. Just like dependency, the constituent approach stems from but is not limited to syntax. The idea of phrase structures influenced many other fields of linguistics as it was, and sometimes still is, the only formal representation taught in the ordinary linguistics curriculum.

The generative grammatical tradition that, in its origins, solely attempts to construct a system that distinguishes grammatical from ungrammatical sentences, left linguistics in a state where the result of the grammatical analysis, namely phrase structure, was difficult to connect to deeper (semantic, conceptual) structures. The result was a far-reaching separation between, on the one side Natural Language Processing (NLP) that needed deeper analyses for parsing, translation, classification, generation etc., and, on the other side, generative linguistics that

built complex structures with the declared goal to model Language as a whole, where the structures got more and more complicated the further the described language diverged from English. In the second half of the 20th century, only a few linguists, often referring themselves to Tesnière, continued to describe language in terms of dependency, mainly because they were working on free word order languages, where the use of phrase structure is more clearly absurd.

Since the 1990s, NLP is turning towards dependency analysis and in the past five years, dependency has become hegemonic. Thus, it is illustrative that the very large majority of parsers presented at recent NLP conferences are explicitly dependency based; Machine Translation is also moving more and more towards dependency, etc. It seems, however, that the connection between computational linguists and dependency linguists remains sporadic: What happens commonly is that someone transfers an existing treebank into a dependency format that fits his or her needs, and other NLP researchers attempt to reproduce this annotation, with statistical or rule-based grammars. Not that the situation was better when parsers still automated phrase structure construction and linguistics discussed "move alpha". Yet, we believe that the state of affairs is different today and dependency linguists and computational linguists have a lot to share.

2.   This volume

The Dependency Linguistics conference Depling 2011 brought together a number of scholars from theoretical and applied linguistics as well as from the domain of Natural Language Processing. All the submissions to the conference were critically reviewed and commented upon by renowned specialists in the field, three to four for each paper. Their comments were an important contribution to the final versions of the papers. This volume unites some of the linguistically oriented articles from the conference (in their revised and extended forms) and gives a general overview over the current state of the art in dependency centered linguistics.

The volume starts out by delimiting the domain of research: Igor Mel'čuk, arguably one of the most prominent dependency linguists of our days, gives in his paper an introduction to dependency and states criteria that allow for the distinction of different levels of dependency – from semantics, over syntax and morphology. In a similar vein, the two following papers in the spirit of the Prague school of dependency linguistics delimit dependency grammar: Jarmila Panevová and Magda Ševčíková pick up the classical question of the separation between grammar and lexicon, and Pavlína Jinová, Lucie Mladová and Jiří Mírovský tackle the question whether dependency can be extended beyond sentence boundaries. Then, Henrik Høeg Müller and Iørn Korzen demonstrate how the Copenhagen

Dependency Treebank has been extended from syntax to morphology and semantics. Dependency treebank creation is also the subject of Kristiina Muhonen and Tanja Purtonen's paper: They exemplify how to create satisfactory annotation for controversial structures such as, for instance, ellipsis using the help of the future users of the treebank.

The other papers in the volume address more specific phenomena in the light of the dependency paradigm. The paper by Orsolya Vincze and Margarita Alonso Ramos explores the problem of the representation of Spanish person names at different levels of linguistic dependency representation.

The next two papers tackle the problem of the representation of coordination in a dependency framework from different angles: Nicolas Mazziotta proposes a new analysis based on data from Old French, and Markus Dickinson and Marwa Ragheb tackle the annotation of coordination on a corpus of learner language.

Eva Maria Duran Eppler analyses exciting data from bilingual code-switching by taking into account a measure that is specific to dependency grammar: The distance between words that are connected by a dependency link. Vered Silber-Varod shows that in some cases those dependency links can go beyond prosodic boundaries of spoken Hebrew. Continuous segments of the dependency tree, called Catenae, are shown by Thomas Groß to be instrumental not only for modeling syntax, but also for the analysis of morphological phenomena.

The concern of Kateřina Rysová's paper is word order in Czech taking into account information structure and semantics. Timothy Osborne addresses the problem of rising, illustrating it mainly on English and German.

Andreas Pankau tackles the astonishing fact of Wh-Copying in German, and Dina El Kassas, finally, looks at the contrary phenomenon in Arabic: Pronoun dropping.

Overall, we can state that this comprehensive and coherent collection of papers covers all fields of contemporary dependency linguistics, ranging from definitional challenges of dependency to concrete analyses of various cross-linguistic phenomena of syntax in its interplay with phonetics, morphology, and semantics, including phenomena for which classical simple phrase-structure based models have proven to be unsatisfactory.

Table of content

Dependency in language

Igor Mel'čuk
Université de Montréal

This paper presents a general overview of the notion of linguistic dependency and of its application in formal modeling of Language. Three major types of dependency are distinguished: semantic, syntactic and morphological dependencies; all cases of their 14 possible combinations in a sentence are examined and illustrated. Each type of dependency is characterized in some detail. For syntactic dependency, three sets of formal criteria are introduced; for morphological dependency, its two major subtypes – agreement and government – are described. The main advantages of syntactic dependency are shown, as well as a case of its insufficiency (in coordination). The place and the role of phrases within dependency framework are touched upon. The so-called Bracketing Paradox is briefly discussed.

1. Introductory remarks

1.1 The task stated

This text does not present new facts or new ideas about known facts. Its goal is to sum up my own experience of more than half a century of work on linguistic dependencies and to better organize the knowledge acquired over this period. It is based on several publications (Mel'čuk 1963, 1974, 1979, 1988, 2002, 2003 and 2009; Mel'čuk & Pertsov 1987), which allows me to skip a detailed explanation of the nature of linguistic dependency. I will also abstain from rigorously pre-senting the necessary notions and formalisms of the Meaning-Text theory (the reader is kindly invited to consult the appropriate titles: e.g. Mel'čuk 1974:31ff, 1981, 1988:43–101, 1997, 2006:4–11, 2012, 2013 and Kahane 2003). Finally, there will be only a dire minimum of references.

My task is three-pronged: (i) to offer the reader a concise overview of what must be known about linguistic dependencies to successfully use them ("Depen-dencies 101"); (ii) to emphasize the advantages of dependencies in linguistic description; (iii) to sketch the place and the use of phrases (≈ constituents), within a strict dependency approach.

But first, a bit of personal experience.

1.2   Some History

I met syntactic dependency for the first time in 1956 while developing a Hungarian-to-Russian machine translation system: Melčuk 1957. Here is an example from this paper: translation of the Hungarian sentence (1a) into Russian – as (1b).

(1)   a.   *A   legtöbb nyelvnek   sok   idegen eredetű*
           the most   language-SG.DAT many foreign "originary"

           szava   van.
           word-SG.NOM.3SG is

    b.   *V   bol'šinstve   jazykov   est' mnogo slov*
           in majority-SG.PREP language-PL.GEN is   many   word-PL.GEN

           inostrannogo   proisxoždenija.
           foreign-NEU.SG.GEN origin-SG.GEN

   'Most languages have many words of foreign origin'.

At least four problems have to be dealt with by a machine translation system to obtain (1b) from (1a):

- The grammatical number of nouns 'language' and 'words': they are singular in Hungarian – because of a quantifier, which in Hungarian requires the singular of the quantified noun, but must be plural in Russian – for the same reason, since the Russian quantifiers require the plural of the noun (with the exception of so-called "small" numerals: 1, 2, 3 and 4).
- The agreement of the adjective INOSTRANNYJ 'foreign' with the noun PROISXOŽDENIE 'origin' in Russian; in Hungarian, adjectives do not agree with nouns.
- The dative of the noun NYELV 'language' in Hungarian, induced by the verb VAN '[there] is', corresponds to the Russian preposition V 'in', induced by the verb EST' '[there] is', equivalent of VAN.
- Word order: some Hungarian words precede their "governors," while their Russian equivalents follow theirs; cf.: Hung. *szava van* ≡ Rus. *est' slova*, Hung. *eredetű szava* ≡ Rus. *slov proisxoždenija*.

I was unable back then, and I am still unable now, to figure out how to formulate the corresponding rules, if the Hungarian sentence is parsed into constituents, that is, supplied with a phrase structure. The constituency approach, borrowed by computational linguists in the ex-USSR from American and British linguistics, was in 1950's the only well-known formal framework, yet I felt very strongly that there was no way you could translate by means of bracketed phrases. And

my attempts to find a proper way to translate a given text lead me to syntactic dependencies.[1]

The above problems can indeed be easily solved by using syntactic dependencies. Let us consider an approximate dependency tree for both sentences in (1):

(2)
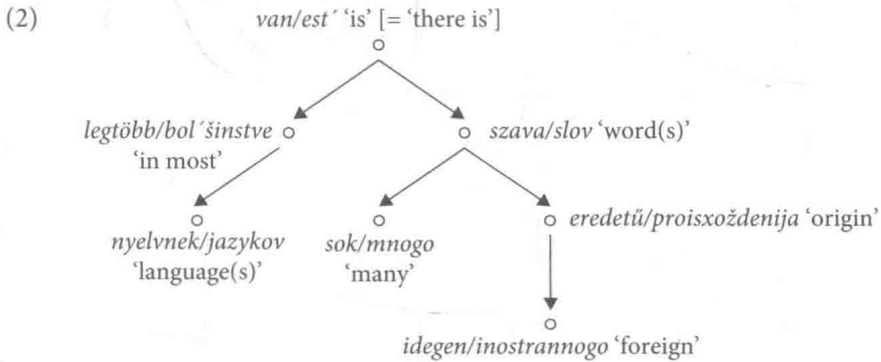
Based on dependencies arrows linking the lexemes, it is obvious how to formulate the rules for the necessary changes between Hungarian and Russian in the four above cases. (Namely: "If there is a quantifier depending on an N, this N is singular in Hungarian and plural in Russian"; "In Russian, an ADJ agrees with the N it depends on"; "In Hungarian, an N depending on the verb VAN as its object must be in the dative, while in Russian, an N depending on the verb EST' as its object must be introduced by the preposition V 'in' and be in the prepositional case"; "An N_1 depending on an N_2 must precede the N_2 in Hungarian, but follow it in Russian."). It became very soon clear that a successful machine translation needs – as a kind of hinge between sentences of two different languages – a syntactic structure, and this structure must be written in terms of dependencies.

1.3   Dependency and Meaning-Text stratificational approach

To see all advantages of dependency representation of utterances, especially in what concerns syntactic dependency, you have to use dependencies in a package together with several other important techniques. I know of three conditions

---

1.   Of course, I was not alone: at least in Germany, France and Czechoslovakia several researchers were inching forward along the same difficult path, and for the same reasons as myself. Interestingly, in the USA, David Hays (1964) and Julia Robinson (1970) formulated explicitly the basic tenets of dependency syntactic description, but theirs remained voices crying out in the desert.

imposed on linguistic research and description that must be met for syntactic dependencies to show all their power.

- Semantic representation as a starting point. The very first thing to do in any linguistic study is to present a formal description of the meaning of the expressions examined – in order to establish the correspondences between the expressions and its meaning. The guiding slogan here is: "We say what we think!"
- Synthetic perspective. A linguistic description is better done from meaning to text: the linguist aims at modeling the activity of the Speaker, who produces texts, not that of the Addressee, who interprets/understands them. The guiding slogan: "To use language is to speak!"
- Stratificational description. Each type of major linguistic unit (sentences and wordforms) is represented under different aspects specific to it, so that different formalisms are needed. As a result, several levels of linguistic representation and different structures within a representation are distinguished; these representations and structures are related by formal rules of the linguistic model. The guiding slogan: "Dead flies and meatballs should be served separately!"[2]

In other words, dependencies and a meaning-to-text stratificational linguistic model have to be used together – lock, stock and barrel.

As far as the characterization of the Meaning-Text model is concerned, two major simplifications are recurred to in this article:

1. While the bottom level is the Semantic representation [= SemR], the upper level in all the examples below is the Deep-Morphological representation [= DMorphR] of the sentence, the reason being that dependencies do not manifest themselves on closer-to-surface levels.
2. Instead of full linguistic representations, I deal here only with their main structures; for instance, instead of the complete SemR of a sentence (which includes the Semantic Structure, the Sem-Communicative Structure, the

---

2. This is a punchline of an old Jewish joke. A poor guy comes to a shabby diner, a typical greasy spoon, and asks for a helping of meatballs. When the dish arrives, he sees several dead flies on the meatballs; calling up the waiter, he indicates the problem to the latter. The waiter explodes in self-assured indignation: – Screw off! If you don't like our meatballs, go somewhere else!! – and starts rolling up the sleeves, getting ready for a physical assault. – No, no, you misunderstood me, – screams the customer. – I have nothing against your meatballs, but I would like to have my dead flies separately and my meatballs separately.

Rhetorical Structure, and the Referential Structure), only its main structure, i.e. the Semantic Structure [= SemS], will be considered.

As far as the proposed definitions of linguistic phenomena are concerned, only prototypical cases of these phenomena are taken into account. Therefore, several definitions given below are incomplete. However, they are sufficient for my purposes here.

## 2.   Different types of linguistic dependency

Let us take a simple sentence and describe it scientifically:

(3)   *Male lions carefully groom their paws.*

The task of the linguist describing a sentence can be reduced to three basic steps:

–   Propose a formal representation of the meaning of this sentence, or its SemR.
–   Propose a formal representation of its physical – that is, phonic – form, or its Phonic representation [= PhonR].
–   Propose a system of rules that, applied to the SemR of sentence (3), produce its PhonR. Such a system, valid for a language as a whole, is a correspondence $\{SemR_i\} \Leftrightarrow \{PhonR_j\}$, or a linguistic model.

I will illustrate this approach by indicating for sentence (3) its SemS and its DMorphS (in conformity with the simplifications formulated in 1.3):

(4)   a.
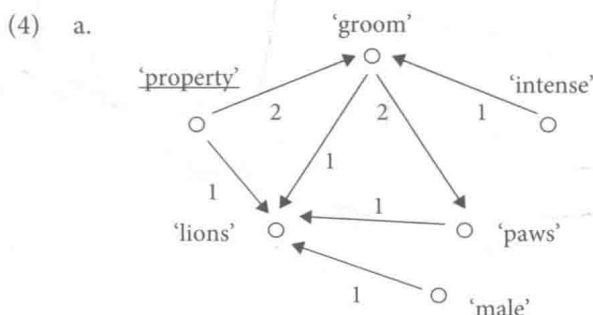
b. $MALE < LION_{PL} < CAREFULLY <$
$GROOM_{IND, PRES, 3, PL} < THEIR < PAW_{PL}$
The symbol "<" indicates linear order.

Literal reading of the SemS in (4a):

'Male lions have the property of intensely grooming their paws'