Linguistic Issues in Language Technology – **LiLT**
Submitted, October 2011

Computational Linguistics without Linguistics? View from Prague

Eva Hajičová

Published by CSLI Publications

LiLT volume 6, issue 6 October 2011

Computational Linguistics without Linguistics? View from Prague

Eva Hajičová, *Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague*

Motto

"My colleagues and I always hoped that linguistics will eventually allow us to strike gold"
Fred Jelinek (in Zampolli Award speech, LREC 2004)

1 Introduction

As noted several times by the pioneers of computational linguistics (such as David G. Hays or Bernard Vauquois), the new field, entering the scientific scene at the beginning of the second half of the last century, originated as a kind of an intersection of already existing well-established scientific fields: linguistics, computer science and cognitive science. During the years, computational linguistics (hereafter, CL) has developed into an independent field having both a scientific and an engineering side (Johnson (2011); perhaps with more emphasis on the engineering, i.e. the natural language processing aspects that Martin Kay and many others would like to see; see e.g. Kay in this volume).

In spite of this develoment, however, computational linguists are sometimes regarded as sitting on two chairs: linguists just say "we do not understand", and therefore they would like to look at computational linguists from a distance and not to integrate them into their (i.e. linguistic) domain, and computer scientists tend to say the same

LiLT Volume 6, Issue 6, October 2011.
Computational Linguistics without Linguistics? View from Prague.

from their perspective and behave in the same way. Also institutionally, some CL institutes or departments or teams are housed in Arts Faculties, and some are affiliated with Computer Science.

We witness a paradoxical situation: Noam Chomsky considered himself a linguist, but he has been largely opposed by traditional linguists. At the same time, his name has been often used by computational linguists affiliated with Schools of Computer Science as an argument that they belong there – Chomsky was referred to as one of the founders of computational science.

Having this situation in mind, it is certainly a timely question to be asked in which way computational linguistics (still?) interacts with linguistics. A quite succint answer has been given by Martin Kay's characterization of CL as a field that tries to do what linguists do in a computational manner (Kay, 2005, p. 429). The aim of the present contribution is to provide some justice to Martin Kay's characterization. We first want to demonstrate on some selected linguistic issues that classical structural and functional linguistics even with its seemingly traditional approaches has something to offer to a formal description of language and its applications in natural language processing (Section 2.1 through 2.4) and to illustrate (in Section 3) by a brief reference to Functional Generative Grammar (on the theoretical side of CL) and Prague Dependency Treebank (on the applicational side) of a possible interaction between linguistics and CL.

2 Some selected principles of structural and functional linguistics

2.1 Introduction

The attributes "structural" and "functional" refer to two important features of the development of linguistics in the first thirty years of the 20th century. These attributes contrast with the orientation prevailing in linguistic studies until then, which were mostly focused on language diachrony, i.e. on historical development of particular languages and language groups. First of all with respect to individual language phenomena, structural linguistics (starting with Ferdinand de Saussure, who found his followers in different linguistic schools of that time) understands language as a system, as a structure of relations. "Functional" then refers to those trends that view language as a functioning system, adapted to its communicative role, and that work with the oppositions of form (*signifiant*) and function (*signifié*).

As follows from the title of the present contribution, the considerations presented here are anchored in the Praguian context, both classical and contemporary, and are based on our own linguistic training and education. However, in spite of this rather personal orientation, we hope that the issues briefly sketched may have some more general impact and consequences. Also, we concentrate only on some selected principles we ourselves consider relevant for the research in computational linguistics and it is quite imaginable that other specialists in the field might feel that some further aspects should have been added.

The principles we focus our attention on are the following: (i) the distinction made between the core of the system on the one hand and its periphery on the other (Section 2.2), (ii) dependency rather than constituency (phrase structure, immediate constituents structure) syntax with an emphasis on the account of underlying syntactic sentence relations (Section 2.3), and (iii) a due respect to the functional aspects of sentence structure (Section 2.4).

2.2 System, core and periphery

When postulating the systematic character of language, structural linguists had soon observed that it is necessary to give up the simple idea of strict compartmentalization of linguistic elements and to regard the classes and subclasses of these elements as formations with a compact, more stable “core” or “center” and with a large “peripheral”, less stable sphere, rather than to look at the system of language as a “system of systems” (subsystems) with clear-cut boundaries. This view had first penetrated to the studies of phonology (in which domain it was elaborated most systematically and in more detail, especially in relation with the theory of markedness) but soon had been understood in a more general sense, being applicable also to language levels higher than phonology and to the system of language as a whole.

The first comprehensive and many-sided “attack” on the issue of core and periphery is contained in the volume *Travaux linguistiques de Prague 2*, with the subtitle *Les problèmes du centre et de la périphérie du système de la langue*, published in Prague in 1966; the topics of the contributions ranged from more general claims (for a more precise reference see Daneš (1966), Leška (1966), or, for a comparison between vagueness and the core-periphery relation, see the paper by Nestupný (1966)) to a discussion of particular issues or linguistic phenomena that reflect in some way the core-periphery relation. All authors, of course, refer back to earlier writings proposing such a view under the same or sometimes different names but analyzing apparently the same issue.

Sgall (2002, 2004, 2009) puts the core-periphery asymmetry into a broader and more complex perspective. He claims that since language is more stable in its core, regularities in language should be searched for first in this core; only then is it possible to penetrate into the subtleties and irregularities of the periphery, the peripheral elements being less stable (they may even totally disappear from the system or on the contrary, perhaps with some modifications, may be shifted to the core, which reflects one aspect of the dynamics of language development). The relatively simple pattern of the core of language (in Sgall's view, not far from the transparent pattern of propositional calculus) makes it possible for children to learn the regularities of their mother tongue. The freedom of language offers space for the flexibility of the periphery.

In the context of computational linguistics, Sgall and Böhmová (2002) touched upon the problem of a possibility to learn from the classical core-periphery opposition also in the CL domain. The authors claim that it is inappropriate to attempt at a specification of both the core and the periphery at once.

This claim of the authors should not be taken as implying that the phenomena should be treated just one by one, separately. The authors propose to aim at a description of the core of the language system by general principles or rules (in which languages differ just in the repertoire of attributes and their values, rather than in the basis of the structural patterns) and to capture the non-prototypical phenomena by rules of a more specific nature. In their opinion, among the marked phenomena there belong e.g. discontinuous constituents (as e.g. the "split" constituent *saw him* in *Him I saw* (see below for the notion of non-projectivity, in Section 2.3) or syntactic relations of another type than dependency (such as e.g. coordination and apposition in: *Charles the Fourth, the King of Bohemia and the Roman Emperor, was one of the greatest figures in Czech history*).

The above attitude can be well reflected both in formal description of language as well as in the build-up of annotated language resources, with an undisputable advantage. In formalizing the general principles or rules, one could reach an account that would come closer to the common human mental capacities and thus to represent a step forward in linking language capabilities with the domain of cognition, without working with a complex innate mechanism specific for the language faculty. In the latter domain, when building an annotation scenario, one should have first in mind the core language patterns, establishing the core categories and subcategories within these patterns and finding then spaces for the account of the peripheral phenomena (be it in the form in a further, more subtle subcategorization, or in the form of enriching the framework by means for some further, non-basic relations such as coordination, or specifying transitions between the core relations and some peripheral, surface-related phenomena as in the case of discontinuous constituents).

To avoid any misunderstanding, it is necessary to stress that the relation of core and periphery has nothing to do with frequency (though it can be observed, as suggested by one of the reviewers, that almost every sentence has something peripheral but that an individual peripheral thing is often relatively scarce in the corpus) and it cannot be related to the distinction of the distribution of linguistic phenomena covered by the notions of "light head" and "heavy tail" (for those notions, see Steedman and also Levin, this volume). Rather, one can say that the core contains relations which make it possible to combine elementary lexical units into unrestrictedly complex sentences; valency (determining the dependency relations that accompany a word) as the meeting point of lexicon and syntax is crucial in this respect (see Section 2.4 below).

2.3 Dependency Syntax

The notion of dependency has been for a long time a matter of continental syntactic theories, introduced there by Tesnière (1959). He viewed the sentence as a hierarchical structure the center of which is the verb; this structure is described on the basis of binary relations between the verbs and their modifiers; one speaks about the valency of the verb. The dependents on the verb are classified into actants and circonstants; in current terms, this classification corresponds to the classification of dependents into arguments and adjuncts. Not only verbs have their valency frames, but also other word classes such as nouns, adjectives etc.

It is sometimes doubted whether the direction of the dependency relation, namely the determination which element of the pair is the governor and which is the dependent in each pair can be reliably stated. In the prototypical case, the main criterion for this distinction can be based on the possibility that, in the endocentric constructions (i.e. constructions in which the distribution of the whole is identical with that of the governing component), the dependent can be absent, not just deleted on the surface. With the exocentric pairs (i.e. constructions where the distribution of the whole does not equal the distribution of any of its elements), for which the above mentioned criterion by itself could not help to find out which element is the governor and which is the dependent, the principle of analogy on the level of parts of speech can be applied: on the basis of the existence of verbs without object (e.g. *to sit, to sleep, ...*) it can be concluded that the verb is the governor also in constructions such as *to find something*, in which none of the members can be deleted. In the same vein, also the subject (Actor) can be understood as a dependent of the verb since there are verbs without a subject (Lat. *Pluit*, in E. *It is raining*, the subject *it* is just a surface filler absent in the sentence structure proper).

The introduction of the notion of a head and the concept of valency brings into the foreground the connection between grammar and lexicon: it reflects the fundamental aspect of the presence of grammatical information in the lexicon. The valency frame is a part of the lexical entry, in which the obligatory and optional syntactic kinds of dependents of the given word (the head) are registered.

Even though the dependency theory belonged to European linguistics rather than to the mainstream syntactic approaches on the other side of the Atlantic, its formalization is due to an American computational linguist David G. Hays (see esp. (Hays, 1964), but it was hinted at already in his paper with K.E. Harper, 1959); an independent formulation was published in 1961 by Russian linguists G.S. Cejtin and L.N. Zasorina.

In the formal account of dependency relations an important role is played by the strongly restrictive condition of projectivity: if a node a depends on b and there is a node c between a and b in the linear ordering, c is subordinated to b (where subordinated means an irreflexive transitive closure of dependency). Apparently, there are many non-projective constructions in the surface shape of the sentences, but they are peripheral in the sense that they concern only some well-defined structures rather than the whole core. It is then a realistic task to attempt to classify the constructions in which the condition of projectivity is not met in the surface shape of the sentence and on this basis to formulate a description meeting the condition as far as the core of language system is concerned and to account by simple and well-defined means also for the cases of superficial non-projectivity.

The leading modern syntactic theories have been based on American linguistic tradition, though the notion of head can already be found in Bloomfield's major work. It was already the analysis of Robinson (1969, 1970) which threw an interesting light on the possibility of a smooth transition between a phrase-based approach to a dependency based one, when she considered the possibilities of finding a formal framework for Fillmore's case grammar. And the development of the originally constituent-based frameworks has indicated that the recognition of the head of a syntactic structure is necessary.

In this context, we have observed two seemingly contradictory tendencies (Hajičová, 2006): (i) the deeper the analysis goes the greater the need for a distinction between the notions of head and modifier (predicate, argument) is felt; (ii) dependency based considerations have gradually and evasively penetrated to the data oriented (i.e. surface-based) statistical models. The first of these observations is supported by the increasing number of semantically oriented studies in which the notion of "head" in one way or another plays an important role (we quote just the names because the relevant references are obvious): the lexico-semantic analysis by J.J. Katz and P.M. Postal when specifying selection restrictions, the distinction between surface constituent structure and the (underlying) functional structure in lexical functional grammar of J. Bresnan and R. Kaplan, the case grammar by C.J. Fillmore motivated by the conviction that Chomskyan account of deep structure is not deep enough to capture the underlying structure of the sentence, and the introduction of the notion of head (and also the consecutive theta theory) in Chomskyan government and binding theory.

A possible explanation may be looked for in the economy and transparency of the dependency based trees; they work without intermediate structures and are more lexically based (see also the arguments of Levin, this volume). In their applications, the data-oriented systems also aim at a representation of the meaning of the surface shapes of sentences (whatever one can understand under "meaning") so that their attention is focused on a most transparent and economic way (avoiding extra nodes for phrases such as NP, VP etc.) from the surface to the depth. And this is the way offered by dependency analysis.

2.4 Functional Aspects

Among the important aspects of functional approaches to language, there belongs the view of language as a functioning system, adapted to its communicative role, and to describe the sentence structure as adapted to its functioning in discourse. It was in this context that the ideas of what is now more generally referred to as "information structure of the sentence" initially appeared: first, clad in a more or less psychological cover (see the two kinds of the so-called progressions of ideas with Weil (1844), namely *marche parallèle* and *progression*), followed by the convictions that such notions as theme and rheme are matters of pragmatics rather than of syntax proper. It was only in the third quarter of the last century that the idea that topic/focus articulation has its significance also for the representation of the meaning of the sentence (claimed already by Sgall much earlier than that, cf. Sgall (1967b, p. 205f)) penetrated into language descriptions of different trends. It is often left unnoticed that actually the split of transformational grammar into the generative and interpretative semantics wings operated with arguments based on semantic differences between sentences that differ – in our understanding – only in their topic focus articulation (this fact, of course, not being recognized by the authors, but a slight reference to the notion of topic can be found in Chomsky's discussion (Chomsky, 1965, p. 224f); as Lakoff (1969) notes, in this context, the influence of Halliday (1967-1968) played its role): see the sentences *Many men read few books* vs. *Few books were read by many men*, or *John talked about many problems to few girls* vs. *John talked to few girls about many problems* adduced by Lakoff, or *Everybody in the room knows at least two languages* against *At least two languages are known by everybody in this room*, discussed already in Chomsky's Syntactic Structures.

One could argue that it is the presence of structures with quantification rather than the topic-focus articulation of the quoted examples that is responsible for the indicated semantic distinction. However, the Praguian writings from the sixties convincingly demonstrate that it is not difficult to find sentences without quantification that exhibit the same phenomenon (in the examples the capitals indicate the intonation centre): *Russian is spoken in SIBERIA* vs. *In Siberia, RUSSIAN is spoken*, or *John works on his dissertation on WEEKENDS* vs. *On weekends, John works on his DISSERTATION.* In Russian linguistics, such examples have been discussed as *Kurit' ZDES'* [lit. Smoke HERE.] vs. *Zdes' KURIT'* [lit. Here SMOKE]. The sentences quoted also document that the difference cannot be ascribed to the active/passive distinction; neither can it be claimed that the word order always plays a decisive role: consider Halliday (1970)'s famous example from a London underground station: *Dogs must be CARRIED.* With the same word order, but with a change in the placement of the intonation centre one gets a certainly unwanted interpretation: *DOGS must be carried* would imply that everybody stepping on the escalator has to carry a dog (in a similar vein to *Carry DOGS.*). A plausible explanation of the semantic difference covering all these cases is to describe them in terms of difference in their information structure.

It directly follows from the above considerations of the semantic relevance of information structure that this phenomenon has to be reflected both in the formal account of language as well as in (at least some, more advanced) applications including the proposals of annotation scenario of large language resources.

In terms of the communicative function of language, an adequate explanation of information structure of the sentence may be based on the relation of aboutness: the speaker communicates something (the Focus of the sentence) about something (the Topic of the sentence), schematically:

> **F(T):** the Focus holds about the Topic
>
> **∼F(T):** negation — (in the prototypical case) the Focus does not hold about the Topic; in a secondary case, the assertion holds about a negative Topic: F(∼T)

The two (semantic) interpretations of (surface) negation can be illustrated by the two readings of the sentence *Bert did not come because he was out of money.* The former, prototypical one, implied e.g. by the question *What about Bert?* can be paraphrased as *I am saying about Bert that he did not come because he was out of money* (with Topic=*Bert* and Focus=*(he) did not come because he was out of money*); the latter, secondary, is implied by the question *Why didn't Bert come?* and can be paraphrased as *I am saying about the fact that Bert did not come that this was caused by the fact that he was out of money* with Topic=*Bert did not come* and Focus=*(because) he was out of money*). Under this latter interpretation, the scope of negation is restricted to the Topic part of the sentence; the assertion triggered (on this reading) by the *because*-clause in Focus is not touched by negation (the reason of Bert's not coming (absence) is ...). However, there is another reading of the above sentence, e.g. if it is followed by: ... *but because he was on his leave of absence.* Under this interpretation, we understand that Bert came, but for some other reason; the sentence can be paraphrased as *I am saying about the fact that Bert came (i.e. about his presence) that it was not because he was out of money but because* ... (with Topic=*Bert came* and Focus=*not because he was out of money*). Under this interpretation, Bert's coming is entailed (belonging to a presupposition of the sentence) and Bert's being out of money is neither entailed nor negated. It may but need not be the case, as the following possible continuations of the sentence indicate: *Bert did not come because he was out of money but because he was on his leave of absence; he lost his purse* (implying he was out of money) contrasted with *Bert did not come because he was out of money but because he was on his leave of absence; he had received his salary just the day before* (implying: he was not out of money). The scope of negation again concerns Focus, schematically: ∼F(T). What is in the scope of negation is neither asserted, nor presupposed; the *because*-clause triggers an allegation (see Hajičová (1984)).

These considerations — in addition to examples of evident semantic differences between sentences such as those quoted above in this section — have led us to the conclusion that TFA undoubtedly is a semantically relevant aspect of the sentence and as such it should be represented at a level of an integrated formal language description capturing the meaning of the sentence. This level can be understood as the "highest" level of the language description viewed from the point of view of the hierarchy from function to form. The inclusion of TFA into this level can serve well as a starting point for connecting this layer with an interpretation in terms of intensional semantics in the one direction and with a description of the morphemic and phonemic means expressing TFA in the other direction.

It then goes without saying that also on the engineering side of CL, in systems that operate with the meaning of the natural language input/output, this basically linguistic aspect of sentence structure has to be taken into account: both in language analysis and synthesis (generation), be it for machine translation systems, question answering systems, advanced information retrieval systems, summarizing, etc., the system should "recognize" / "formulate" sentences with an appropriate indication e.g. of the scope of negation or the scope of quantifiers. For instance, if the knowledge base contains a piece of information derived from the input sentence *Bert did not come because he was out of money.* it should "know" which interpretation is assigned to it; otherwise, if it answers *Yes, he was* to the incoming question *Was Bert out of money?* this answer might be false in case the interpretation of the incoming information would have been *I am saying about the fact that Bert came (i.e. about his presence) that it was not because he was out of money but because ...* (with Topic=*Bert came* and Focus=*not because he was out of money*). Under this interpretation, Bert's being out of money is neither entailed nor negated. On the other hand, if the sentence in the input were *Because he was out of money Bert did not come*, the reason for Bert's absence is in the topic of the sentence, and as such (non-negated) it is out of the scope of negation. Then the answer *Yes, he was* would be true and fully appropriate. Similar considerations hold for language analysis and generation in machine translation systems.

3 Interaction between linguistics and computational linguistics: Functional Generative Grammar (in theory) and Prague Dependency Treebank (in application)

3.1 Functional Generative Description

Based on the Praguian linguistic tenets, Functional Generative Description (hereafter, FGD) was formulated (see Sgall (1964, 1967b,a), Sgall et al. (1969, 1986)) as an alternative approach to a formal description of language: at the time of its origin (as early as 1963-1964), the only generative description was Chomsky's transformational generative grammar. The attribute "alternative" refers to five basic features: (i) FGD followed the "stratificational avenue for the extension of syntactic models" (as characterized by Hays (1964), comparing it with the transformational one; in this aspect FGD is close to Lamb's stratificational grammar); in the later elaboration of the framework, there remained only two levels, namely the morphemic and the underlying syntactic level called tectogrammatical; (ii) it did not work with transformations; (iii) it introduced the notion of dependency relation into the description of the syntactic structure; originally, the rules were close to phrase structure rules with the governing element being marked, and later, from 1969 on, a fully dependency description of syntax has been postulated and used; (iv) the awareness that a due regard is to be paid to the functions of language (which is reflected in FGD since as early as 1967 by integrating the information structure into the formal description of the tectogrammatical level); and (v) a due regard to the distinction between linguistic meaning (belonging to the description of the system of language) and cognitive (extralinguistic) content.

This theoretical model works with an underlying syntactic level called tectogrammatics, which is understood as the interface level connecting the system of language (cf. de Saussure's notion of langue as an abstract structure of relations and properties used by the speaker to produce concrete utterances and by the hearer to understand them; the similar notion of linguistic competence as coined by Chomsky covers not only an inventory of units but also as a system of rules for the generation of utterances) with the cognitive layer, which is not directly mirrored by natural languages. Language is understood as a system of oppositions, with the distinction between their prototypical (primary) and peripheral (secondary, marked) members. We assume that the tectogrammatical representations of sentences can be captured as dependency based structures the core of which is determined by the valency of the verb and of other parts of speech. Syntactic dependency is handled as a set of relations between head words and their modifications (arguments and adjuncts). However, there are also the relations of coordination (conjunction, disjunction and other) and of apposition, which we understand as relations of a "further dimension". Thus, the tectogrammatical representations are more complex than mere dependency trees.

The core of a tectogrammatical representation is a dependency tree the root of which is the main verb. Its valency is understood as the set of dependency relations (called functors in FGD) between the verb (head, governor) and the items dependent on that verb. The dependent members are divided into arguments (i.e. inner participants) and adjuncts (circumstantials or free modifications); in the more recent treatments, an intermediate class of modifications of the verb has been characterized (Panevová, 2003) which shares some of the features of arguments and some of adjuncts.

The (underlying) subject is understood as one of the participants, although it has certain specific properties, being in a sense more loosely connected with the verb than other dependents are. FGD works with five arguments (Actor/Bearer, Addressee, Patient, Origin and Effect). Among the typical adjuncts there are Locative, several Directional and Temporal modifications, Condition, Means, Manner, etc. If the valency frame of a verb contains only a single participant, then this participant is its Actor, even though (in marked cases) it corresponds to a cognitive item that primarily is expressed by some other participant. Also nouns, adjectives and some other word classes have their own valency.

In a tectogrammatical representation, there are no nodes corresponding to the function words (or to grammatical morphs). Correlates of these items (especially of prepositions and function verbs) are present there only as indices of node labels: the syntactic functions of the nodes (arguments and adjuncts) are rendered as functors and subfunctors, and the values of their morphological categories (tense, number, and so on) have the forms of grammatemes.

Dependency trees on the tectogrammatical layer are projective (unimportant exceptions aside), i.e. for every pair of nodes in which a is a rightside (leftside) daughter of b, every node c that is less (more) dynamic than a and more (less) dynamic than b depends directly or indirectly on b (where indirectly refers to the transitive closure of depend). This strong condition together with similar conditions holding for the relationship between dependency, coordination and apposition, makes it possible to capture the tectogrammatical representations in a linearized way, by a parenthesized string. Projective trees thus come relatively close to linear strings; they do not surpass the generative capacity of context free grammars and can be adequately represented by bracketed strings (with every dependent being enclosed in its own pair of brackets).

Dependency based representations make a rather straightforward description of the information structure of the sentence (its topic-focus articulation, TFA in the sequel) possible. The tectogrammatical representations reflect the topic-focus articulation (information structure) of sentence including the scale of communicative dynamism (underlying word order) based on the dichotomy of contextually bound (cb) and non-bound (nb) items: for every autosemantic lexical item in a sentence (i.e. for every node of its tectogrammatical representation) it is specified whether it is (a) contextually bound (cb), i.e. an item presented by the speaker as referring to an entity assumed to be easily accessible by the hearer(s), more or less predictable, readily available to the hearers in their memory, or (b) contextually non-bound (nb), i.e. an item presented as not directly available in the given context, as cognitively "new". While the characteristics "given" and "new" refer only to the cognitive background of the distinction of contextual boundness, the distinction itself is an opposition understood as a grammatically patterned feature, rather than in the literal sense of the term. This point can be illustrated by the sentence *My mother recognized only HIM, but none of his FRIENDS.* In the context such as *Yesterday, we were visited by Tom and his friends.* Both Tom and his friends are "given" by the preceding context, but their linguistic counterparts are structured in the given sentence as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center indicated here by capitals).

In the prototypical case, the head verb of the sentence and its immediate dependents (arguments and adjuncts) constitute the Topic of the sentence if they are contextually bound, whereas the Focus consists of the contextually non-bound items in such structural positions (and of the items syntactically subordinated to them). Also the semantically relevant scopes of focus sensitive operators such as *only*, *even*, etc. can be characterized in this way (for a discussion concerning the complexity and the possibilities to formulate a formal semantic account of this phenomenon, see also Hajičová et al. (1998)).

There are two reasons to distinguish the opposition of contextual boundness as a primary (primitive) one and to derive the Topic-Focus bipartition from it. First, and most importantly, the Topic/Focus distinction exhibits — from a certain viewpoint — some recursive properties, exemplified first of all in sentences which contain embedded (dependent) clauses. The dependent clause D functions as a sentence part of the clause containing the word on which D depends, so that the whole structure has a recursive character; one of the questions discussed is whether the T-F articulation should be understood as recursive, too. Several situations arise: (i) one of the clauses may be understood as the F of the whole sentence, though each of the clauses displays a T-F articulation of its own; (ii) in the general case the boundary between T and F may be within one of the clauses. Thus in the sentence (which is a translation of a Czech sentence in the Prague Dependency Treebank) *While the market with radio signal is* *already saturated, // unexploited possibilities still exist within regional and local transmission of the television signal.* In the given context (implying a saturation of the radio-signal market) the boundary between (global) topic and (global) focus of the whole complex sentence is indicated by the double slash (//); the local foci of the individual clauses are marked by underlining. The sentence *Our younger colleagues, who recently finished their doctoral studies, // compete for scholarships abroad.* is an example of the general case: here again, the boundary between the (global) topic and (global) focus is indicated by a double slash, the local focus of the embedded clause being indicated by underlining.

The second argument is related to the fact that Topic/Focus bipartition cannot be drawn on the basis of an articulation of the sentence into constituents but requires a more subtle treatment. In early discussions on the integration of the topic-focus articulation into a formal description of grammar, the proponents intended to specify this aspect of the structure of the sentence in terms of the type of formal description they subscribed to. Within the framework of generative transformational grammar, Chomsky (1971, p.205) defined focus as "a phrase containing the intonation center", i.e. in terms of constituency (phrase-structure) based description (see also Jackendoff (1972, p.237)). Such a description served as a basis also for several studies on the relationship between syntax and prosody (e.g. Schmerling (1976), Selkirk (1984)): the boundaries between topic and focus or some more subtle divisions were always supposed to coincide with the boundaries of phrases.

However, the definition of Focus (and of presupposition, in Chomskyan terms) as a phrase is untenable since it is not always possible to assign the focus value to a part of the sentence that constitutes a phrase. This claim is supported by examples such as *John went for a week to Sicily. (He didn't go only for a weekend to his parents.)*; in the context indicated by the continuation in the brackets, the Focus of the sentence is *for a week to Sicily*, which would hardly be specified as constituent under the standard understanding of this notion. It was convincingly argued by Steedman (1996, 2000) that it is advisable to postulate a common structure for accounting both for the syntactic structure of the sentence, as well as for its information structure. For that purpose, Steedman proposes a modification of categorial grammar, called combinatory categorial grammar. A syntactic description of a sentence ambiguous in information structure should be flexible enough to make it possible to draw the division line between Topic and Focus also in other places than those delimiting phrases; in Steedman (1996, p.5), the author claims that e.g. for the sentence *Chapman says* *he will give a policeman a flower* his "theory works by treating strings like *Chapman says he will give, give a policeman* and *a policeman a flower* as grammatical constituents" and thus defining "a constituent" in a way that is different from the "conventional linguistic wisdom".

The representation of such an ambiguity in a dependency framework like that of the Praguian Functional Generative Description causes no difficulty. In case the root of the tree (the verb) is cb, then it depends on the cb/nb feature of its dependents whether *Chapman says* or *Chapman says he will give, says will give a policeman a flower*, or *a policeman a flower* are the elements of the Topic, answering the question *What does Chapman say*, or *What does Chapman say he will give whom?*, or *Who says he will give a policeman a flower*, respectively (in the last context, the spoken form of the sentence would have the intonation center on *Chapman*). If the verb is nb, then again different divisions are possible: either the whole sentence is the Focus (What happened?), or the verb and some of the dependent nb elements are elements of the Focus. In the underlying tree structure, the cb nodes depend on the verb from the left, the nb nodes from the right. A division line between Topic and Focus can then be characterized as intersecting an edge between a governor and its dependent (the latter may be a single node or a subtree), with the provision that whatever is to the right of the given dependent in the tectogrammatical dependency tree, belongs to the Focus, the rest to the Topic.

3.2 The Prague Dependency Treebank

The Prague Dependency Treebank (PDT; for an overall characterization see e.g. Hajič (1998)) is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus (CNK), with a mark-up on three layers: (a) morphemic, (b) surface shape "analytical", and (c) underlying (tectogrammatical). The current version (the description of which is publicly available at `http://ufal.mff.cuni.cz/pdt2.0`, with the data itself available from the LDC as catalog No. LDC2006T01), annotated on all three layers, contains 3165 documents (text segments mainly of a journalistic genre) comprising of 49431 sentences and 833195 occurrences of tokens (word forms and punctuation marks).

On the tectogrammatical layer, which is our main concern from the theoretical point of view, every node of the tectogrammatical representation (TGTS, a dependency tree) is assigned a label consisting of: the lexical value of the word, its "(morphological) grammatemes" (i.e. the values of morphological categories), its "functors" (with a more subtle differentiation of syntactic relations by means of subfunctors, e.g. "in", "at", "on", "under"), and the topic-focus articulation (TFA) attribute containing values for contextual boundness (for a motivation for the introduction of this value see Section 3.1 above). In addition, some basic coreferential links (including intersentential ones) are also added. It should be noted that TGTSs may contain nodes not present in the morphemic form of the sentence in case of surface deletions.

The tectogrammatical tree structures are projective. In the annotation of PDT, we work also with (surface) analytic representation, a useful auxiliary layer from the technical point of view, on which the dependency trees include nodes representing the function words and the tree reflects the surface word order. This combination allows for non-projective structures in cases such as *A neighbour came in, who told us this* (with the relative clause dependent on the subject noun). We assume that such cases can be described as surface deviations from the underlying word order (i.e. in a tectogrammatical representation corresponding to the example given above, the main verb is not placed between the subject and the dependent clause).

4 Concluding Remarks

In conclusion of our reflections on the relation between linguistics and computational linguistics, we would like to make two additional remarks.

First, not only the relation going from linguistic theory to computational models and their applications is important but also the opposite direction should be mentioned. Any modern linguistic theory has to be formulated in a way that it can be tested by some testable means. One of the ways to test a theory is to use it as a basis for a consistent annotation of large language resources, i.e. of text corpora. Annotation may concern not only the surface and morphemic shapes of sentences, but also (and first of all) the underlying sentence structure, which elucidates phenomena hidden on the surface although unavoidable for the representation of the meaning and functioning of the sentence, for modelling its comprehension and for studying its semantico-pragmatic interpretation. One of the aims the PDT was designed for was to use it as a testbed for the theoretical assumptions encapsulated in the Functional Generative Description.

Second, the motto of my present contribution refers to the Zampolli Award speech given by the pioneer in statistical methods in CL Frederick Jelinek at the LREC conference in 2004 in Lisbon. It indicates among other things that problems that lead to using plain text data for statistical modeling and learning are only a minority in the portfolio of important computational linguistics problems; in the majority of problems more or less profound linguistic expertise to at least divide the problem into manageable modules is needed: in machine translation there is the need to lemmatize and to identify at least phrases or dependency pairs, for predicate extraction at least some syntactic structure of the sentence is needed, for word sense disambiguation it should be clear which senses of the given words are to be disambiguated, and even speech recognition needs transcripts of what has been said, or what can be reconstructed from the speech signals. A parallel can be drawn between an "observation" and "measurement" in physics or astronomy on the one hand and in linguistics on the other; however, there are no objective criteria for anything but plain text corpora and digitalized speech recordings and an intermediary is necessary, which is a human, with her or his intuition about the relationship between the observed data (text or speech) and their meaning through her or his understanding and interpretation, capable of formalizing his or her interpretation and understanding. This is the role of the annotators and the results of their work are the annotated linguistic data. Using part of the annotated data for machine learning is a step further: we get rid of the ineffective AI-style estimation of various weights, preferences and such factors, which apparently can be more effectively estimated by the computer. When touching upon the issue of (merely) an apparent difference between the grammar-based (rule-based) and statistical parsers with respect to their linguistic background and orientation, Johnson (2009) notices that corpus annotation plus statistical inference seems to be a more effective way of getting linguistic information into a computer than manually writing a grammar (see also Johnson, this volume). In our opinion, one of the reasons for this is the fact that with manual writing of rules, one takes one linguistic phenomenon after another, which is a lengthy, time-consuming procedure and may lead to disregarding the interaction of phenomena; when annotating a corpus, the annotators also tackle one occurrence of a phenomenon after another but the statistical inference applies a more global view taking care of all the possible (and captured) analyses.

Thus the answer to the question *Is there any place today for linguistics and linguists (apart from annotation) in CL?* is a definite YES. To take just a rather fashionable example of today, machine learning: especially for complex language applications, machine learning cannot be effectively solved without a prescribed and relatively fixed model structure. For example, for an effective part-of-speech tagging such issues have to be tackled as which features are important: is it the word to left or the word to the right? How far can we go in the context? Should we use the tags assigned to the neighboring words in the process? And what information should be included in the tags? As Moore (2009) has put it, knowledge of linguistic phenomena leads to understanding the limitations of particular statistical models and to better feature selection for such models. The machine learner cannot be instructed "please use any context and any combination of features and tell me which are important" — there are simply too many of them. It is here where the statistician and programmer should really talk to the linguist(s) and come up with appropriate characteristics.

Acknowledgments

The present contribution is partly based on my previous papers on the topic, namely on my talks at ACL Meeting in Sydney in 2006 and at the Computational Linguistics Conference in 2008 in Haifa (Hajičová, 2006, 2008). The author gratefully acknowledges the financial support of the research grant of the Ministry of Education of the Czech Republic MSM 0021620838. Many thanks are due to Jarmila Panevová, Petr Sgall and Jan Hajič, who have read the preliminary draft of the paper and offered me several most useful comments and hints, and to the two anonymous reviewers most of whose suggestions are reflected in the final wording of the paper.

References

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge (Mass.): The M.I.T. Press.

Chomsky, Noam. 1971. On generative semantics. In D. D. Steinberg and L. A. Jakobovits, eds., *Semantics*, pages 183–216. Cambridge (Mass.): The M.I.T. Press.

Daneš, František. 1966. The relation of centre and periphery as a language universal. *Travaux linguistiques de Prague 2, Les problèmes du centre et de la périphérie du système de la langue* pages 9–21.

Hajič, Jan. 1998. Building a syntactically annotated corpus: The prague dependency treebank. In E. Hajičová, ed., *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum: Prague.

Hajičová, Eva. 1984. Presuppositions and Allegation Revisited. *Journal of Pragmatics* 8:155–167.

Hajičová, Eva. 2006. Old linguists never die, they only get obligatorily deleted. *Computational Linguistics* 32(4):457–469.

Hajičová, Eva. 2008. What are we talking about and what we are saying about it. *Computational Linguistics and Intelligent Text Processing* 4919/2008:241–262.

Hajičová, Eva, Barbara H. Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Dordrecht: Kluwer.

Halliday, M. A. K. 1967-1968. Notes on transitivity and theme in English. *Journal of Linguistics* 3 and 4:37–81, 199–244, 179–215.

Halliday, M. A. K. 1970. *A Course in Spoken English: Intonation*. Oxford: Oxford University Press.

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

Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Johnson, Mark. 2009. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL Workshop on the Interaction between Linguistics and Computational Linguistics*, pages 3–11. Athens, Greece.

Johnson, Mark. 2011. How relevant is linguistics to computational linguistics. *Linguistic Issues in Language Technology (LiLT)* this volume.

Kay, Martin. 2005. A Life of Language. *Computational Linguistics* 31:425–438.

Lakoff, George. 1969. On generative semantics. In D. D. Steinberg and L. A. Jakobovits, eds., *Semantics*, pages 232–296. Cambridge (Mass.): The M.I.T. Press.

Leška, Oldřich. 1966. "Le centre" et "la périphérie" des différents niveaux de la structure linguistique. *Travaux linguistiques de Prague 2, Les problèmes du centre et de la périphérie du système de la langue* pages 53–57.

Moore, Robert C. 2009. What do computational linguists need to know about linguistics. In *Proceedings of the EACL Workshop on the Interaction between Linguistics and Computational Linguistics*, pages 42–52. Athens, Greece.

Nestupný, Jiři V. 1966. On the analysis of linguistic vagueness. *Travaux linguistiques de Prague 2, Les problèmes du centre et de la périphérie du système de la langue* pages 39–51.

Panevová, Jarmila. 2003. Some Issues of Syntax and Semantics of Verbal Modificatios. *Proceedings of MTT 2003, First International Conference on Meaning-Text Theory* pages 139–146.

Robinson, Jane J. 1969. Case, caegory and configuration. *Journal of Linguistics* 6:57–80.

Robinson, Jane J. 1970. Dependency structures and transformational rules. *Language* 46:259–285.

Schmerling, S. F. 1976. *Aspects of English sentence stress*. Austin, Texas: University of Texas Press.

Selkirk, E. O. 1984. *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.

Sgall, Petr. 1964. Generative Beschreibung und die Ebenen des Sprachsystems. In *Zeichen und System der Sprache III, Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung, Nr. 11*, pages 225–239. Berlin.

Sgall, Petr. 1967a. Functional Sentence Perspective in a Generative Description. *Prague Studies in Mathematical Linguistics* 2:203–225.

Sgall, Petr. 1967b. *Generativni popis jazyka a česká deklinace (Generative Description of Czech and Czech Declension.* Prague: Academia.

Sgall, Petr. 2002. Freedom of language: Its nature, its sources and its consequences. In *Prague Linguistic Circle Papers 4*, pages 309–329. Benjamins: Amsterdam/Philadelphia.

Sgall, Petr. 2004. Types of language and the simple pattern of the core of language. In P. Steckenburg, ed., *Linguistics Today – Facing a Greater Challenge (Plenary lextures from the 17th Congress of Linguists)*, pages 22–43. Benjamins: Amsterdam/Philadelphia.

Sgall, Petr. 2009. Where to look for the fundamentals of language. *Linguistica Pragnesia XIX/1* pages 1–35.

Sgall, Petr and Alena Böhmová. 2002. The simple core and the complex periphery of natural language a formal and a computational view. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects.* Dordrecht and Prague: Reidel and Academia.

Sgall, Petr, Ladislav Nebeský, Alla Goralčiková, and Eva Hajičová. 1969. *A Functional Approach to Syntax in Generative Description of Language.* New York: American Elsevier.

Steedman, Mark. 1996. *Surface Structure and Interpretation.* Cambridge, MA: MIT Press.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry* 31:649–689.

Tesnière, Lucien. 1959. *Eléments de syntaxe structural.* Paris: Klinksieck.

Weil, Henri. 1844. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes. (Translated as The Order of Words in the Ancient Languages Compared with That of the Modern Languages, Boston 1887, re-edited Amsterdam 1978).* Paris.