Resemblances between Meaning⇔Text Theory and Functional Generative Description

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics, Charles University,
Malostranské nám. 25, CZ-11800 Prague, Czech Republic
zabokrtsky@ckl.mff.cuni.cz

Abstract.
Functional Generative Description (FGD) is a stratificational dependency-based approach to natural language description, which has been developed by Petr Sgall and his collaborators in Prague since 1960's. Although FGD bears surprisingly many resemblances with the Meaning-Text Theory, to our knowledge there is no reasonably detailed comparative study available so far. In this paper, we try to point out at least the basic obvious parallels (and also differences), which – in our opinion – remain generally unknown.

1 Introduction

Nowadays, one can find a few comparisons of MTT and Chomskian streams in the literature, for instance [Nakhimovsky1990] or [Bolshakov and Gelbukh2000], but hardly anything has been written about the relation between MTT and FGD, although they have so much in common. Just for the beginning:

- both approaches originated almost at the same time ([Žolkovskij and Mel'čuk1965] for MTT, [Sgall1967] for FGD);
- both approaches share the same roots (or, at least, the similar background) in European structural linguistics;
- both approaches are stratificational;
- both approaches are dependency-oriented – they use dependency trees as the representation backbone on both syntactic levels;
- although both approaches proved to be useful and successful in large-scale implementations (machine translation system ETAP for MTT, [Apresian et al.2003]; Prague Dependency Treebank (PDT)[1] for FGD, [Hajič et al.2001]), they are still outside the main stream of computational linguistics, which is governed by phrase-structure models.

We are not aware of any larger text comparing MTT and FGD – in the literature we found only a few marginal remarks (e.g. in [Sgall1997]). We believe that there is hardly any third approach sharing so much with MTT and that is why we would like to draw attention to selected similarities between MTT and FGD. In the presented comparison (which is still preliminary and far from exhaustive) we will compare MTT and FGD from the following viewpoints:

- in Section 2 we show the similarity of FGD and MTT using the list of MTT features enumerated in [Bolshakov and Gelbukh2000];
- in Section 3 we compare the systems of levels used in MTT and FGD;
- in Section 4 we compare deep-syntactic level of MTT with tectogrammatical level of FGD.

[1] FGD served as the theoretical framework for PDT. However, there are many subtle differences between the annotation scheme of PDT and the original FGD.

2 "Peculiarities of the MTM"

We borrowed the title of this section from [Bolshakov and Gelbukh2000], where the authors present some properties of MTT as "peculiarities of the MTM [Meaning-Text Model], which it retains as compared with other linguistic models". In the following paragraphs, we go through most of these properties and show that in fact they could be ascribed to FGD too:

"Orientation to synthesis." Both approaches put originally emphasis on language synthesis (generation) rather than on analysis. However, in their implementations they proved to be able to work also in the opposite direction (building trees above a raw corpus in case of PDT, analyzing a sentence in the input language in case of ETAP)

"Multilevel character of the model." In both approaches, language description is organized into a set of linearly ordered levels (Section 3).

"Distinguishing deep and surface syntactical representations." FGD distinguishes deep and surface syntactical levels in a very similar way to MTT.

"Independence between the composition of words and their order in a sentence." As in other dependency formalisms, FGD clearly distinguishes between dependency and linear ordering of words. The nodes of the dependency trees at the tectogrammatical level of FGD are still ordered, but this ordering is used for representation of information structure instead of surface order.

"Accounting of communicative structure of text." Communicative structure (or information structure, in the Praguian terminology) is a part of the tectogrammatical level of FGD.

"Orientation to languages of a type different from English.". FGD was originally developed for Czech, but English was often used for illustration of some aspects too. Recently, tectogrammatical representation of English is being elaborated in detail ([Cinková2004]) and also experiments with tectogrammatical representation of German have been performed; the layered annotation scenario of PDT (morphological, analytical and tectogrammatical level) derived from FGD was also used for building a treebank of Arabic ([Hajič et al.2004]).

"Labeling syntactic relations between words." Again, similarly as other dependency formalism, FGD does label the arcs of dependency trees (on the tectogrammatical level of FGD, these labels are called functors).

"Keeping traditions and terminology of classical linguistics." In this aspect, FGD obviously followed the tradition of the Prague School.

3 System of levels

Both in MTT and FGD, the relation between meaning and text/sound is defined via a sequence of correspondences between adjacent levels ([Sgall et al.1986], page 7: 'among others, the Prague School elaborated it [the distinction between the meaning and sound patterns] as a sequence of well-motivated levels articulating the complex relation between meaning and sound.'). In Praguian perspective, such correspondences are interpreted as a relation between form and function.

There are seven levels in MTT and five levels in the original FGD. Their rough alignment depicted in Figure 1 should mirror the following observations:

- In FGD, there is no direct counterpart of what is called semantic representation in MTT.[2] No notions comparable the notions of lexical function or semantic primitives were elaborated within FGD so far. However, there is a certain overlap of the MTT-semantic level with FGD-tectogrammatical level, namely they both contain the representation of the information structure (resp. communicative structure).

[2]However, the necessity of a higher level was declared in FGD from the very beginning, see [Sgall1967], p. 56.

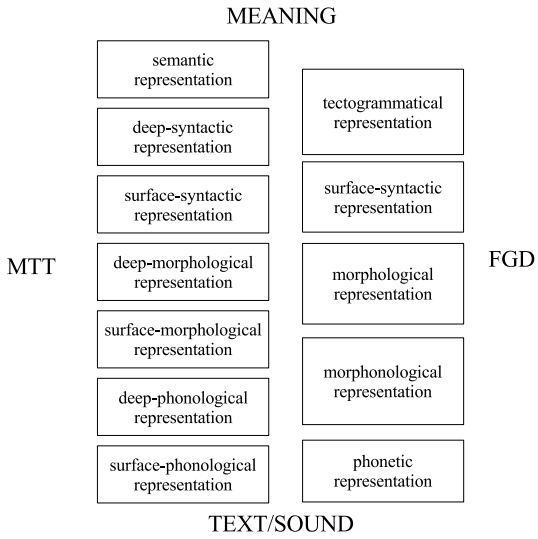

Figure 1: Levels of utterance representation in MTT and in FGD.

- tectogrammatical level matches almost perfectly the deep syntactic level of MTT, some details will be discussed in 4. But again, there is a certain overlap between surface syntactic level of FGD and deep syntactic level of MTT, because deep-syntactic relations in MTT ([Mel'čuk1988], pp. 64-65) seem to be more surface-oriented than functors in tectogrammatical level.
- Surface syntactic level of MTT almost perfectly matches surface syntactic level in PDT:[3] both representations have a form of a dependency tree where nodes correspond to the individual words of the sentence and the arcs are labeled with names of surface syntactic relations.

As the time went by, purely probabilistic models rather than linguistically motivated solutions were more successful in the field of spoken language processing, which is probably the reason why modern implementations of both approaches do not focus on the lowest levels of language description any more. We will not make any comments on the lower levels here.

As for the highest levels, both approaches state in a similar way where their understanding of (linguistic) meaning ends:

- [Mel'čuk1988], p. 44: "'Meaning' ... refers only to information conveyed by language; in other words, it is whatever can be extracted from or put into an utterance solely on the basis of linguistic skills, without recourse to encyclopedic knowledge, logic, pragmatics or other extralinguistic abilities", p. 46 : "... the correspondence between semantic representations and cognitive representations (...) goes beyond the scope of MTT"
- [Sgall et al.1986], p. 13: "the concept of meaning (...) has to be distinguished from that of (cognitive, ontological) content (or factual knowledge)... The analysis of meaning (...) cannot be replaced by an analysis of 'extralinguistic' reality itself."

Their attitude to semantically or logically defective sentences also seems to be the same (perhaps in allusion to the discussion around Chomsky's *sleeping green ideas*):

[3]We have not found an explicit description of the FGD surface syntax level in literature, that is why we compare MTT only to PDT at this point. Moreover, in the recent versions of FGD (roughly since 1990's), surface syntax is not treated as an autonomous level of language description.

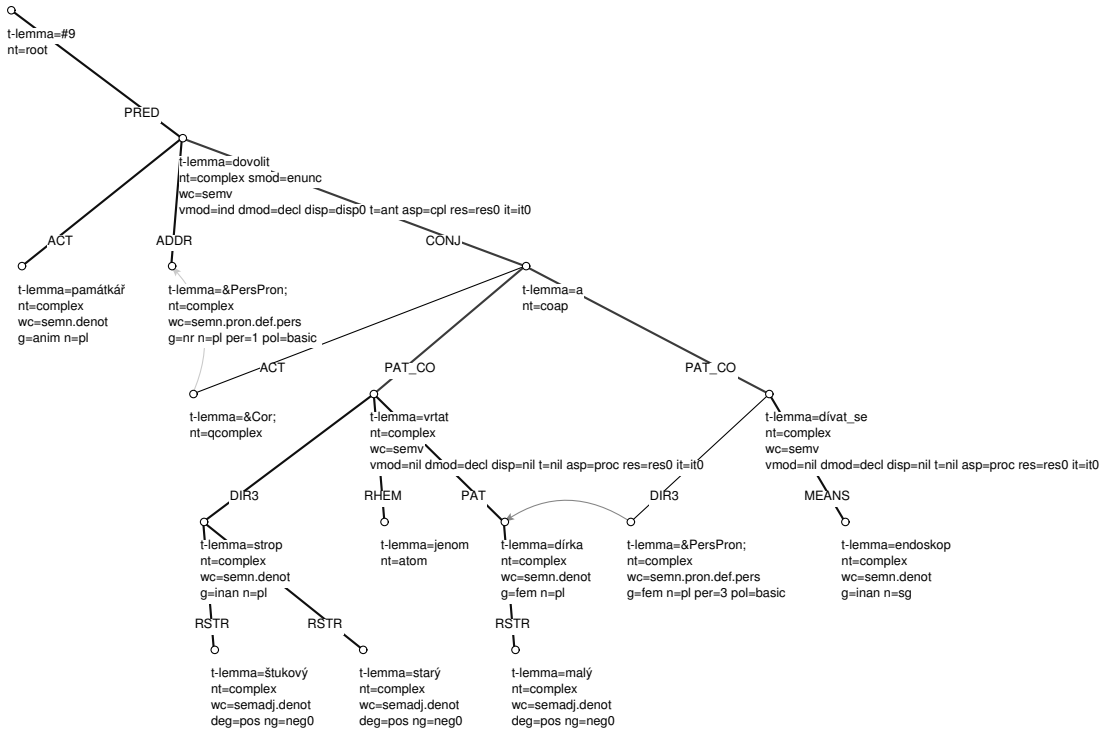

Figure 2: Simplified tectogrammatical tree structure of the sentence "*Do starých štukových stropů nám památkáři dovolili vrtat jenom malé dírky a dívat se do nich endoskopem.*" from PDT 2.0. (The conservationists allowed us to drill only small holes into the old parget ceilings and to look into them with an endoscope.)

- [Sgall et al.1986], p. 106: "A sentence meeting the restrictions of strict subcategorization, but not the selectional restrictions, seems to have a meaning (or even several): witness the fact that it can be translated into other languages... Thus, it is not advisable to 'asterisk away' sentences that merely haven't found any occasion of use, thanks to our image of the world."
- [Mel'čuk1988], p. 47: - "the analysis of meaning itself goes beyond the scope of MTT: it does not distinguish "normal" meanings from absurdities, contradictions or trivialities. Discovering that something is stupid or absurd or detecting contradictions is by no means a linguistic task."

4 Deep-syntactic level vs. Tectogrammatical level

In the following list we would like to note several similarities between the tectogrammatical level in FGD and deep-syntactic level in MTT (a simplified sample of tectogrammatical tree representation is depicted in Figure 2; the example comes from Prague Dependency Treebank 2.0,[4] which will be released in the near future):

- the skelet of both representations is formed by dependency tree (unordered in MTT, ordered according to information structure in FGD);
- only semantically full lexemes (autosemantic words) do have nodes of their own (semantically empty lexemes/synsemantic words, such as prepositions, subordinating conjunctions,

[4]See `http://ufal.mff.cuni.cz/pdt2.0`

auxiliary verbs etc. are introduced only in the surface-syntactic structure);

- each lexeme is associated with appropriate semantically full grammemes (grammatemes in FGD terminology); grammemes imposed only by government and agreement are excluded;
- both approaches work with fictitious lexemes (needed e.g. to encode semantically non-empty syntactic constructions);
- dependency tree is accompanied with (non-tree) grammatical coreferential relations, together forming dag (directed acyclic graph).[5]

Occasionally, the question of convergence of the Western main stream (computational) linguistics with the dependency-based approaches is discussed. In our opinion, it can be nicely illustrated on the recent developments in the field of treebanking. In [Kingsbury and Palmer2002], a layer of predicate-argument structures was added to English Penn Treebank – they started by marking clause nuclei composed of verbal predicates and their arguments. In fact, these structures can already be viewed as small (still rather flat and isolated) dependency trees, headed by (potentially complex) verbal forms. Later, they added also 'modifiers of event variables', thus broadening the nuclei of dependency trees with what FGD would call free modifiers (e.g. [Babko-Malaya et al.2004]). Then also the argument structures for instances of common nouns were added ([Meyers et al.2004]). Finally, the isolated islands containing small dependency trees were connected with relations corresponding to subordinate and coordinating conjunctions [Miltsakaki et al.2004], thus forming a deeper structure, extremely similar to deep-syntactic (resp. tectogrammatical) dependency trees available in MTT and FGD decades ago.

5 Final Remarks

Besides many similarities between FGD and MTT, one can find also several points in which the two approaches differ:

- Compared to MTT, FGD invested less energy into lexicologic and lexicographic issues. Although for instance an electronic valency lexicon based on FGD is being intensively built now ([Straňáková-Lopatková and Žabokrtský2002]), no studies fully comparable to [Mel'čuk and Žolkovskij1984] were created in the context of FGD.
- MTT and FGD elaborated valency theories (e.g. [Panevová1980] for FGD), but in spite of the fact that they both refer back to [Tesnière1959], their criteria for distinguishing *actants* from *modifiers* substantially differ.
- Both approaches prefer different representation of coordination constructions – FGD represents coordination by adding a new dimension to the tectogrammatical tree (by using different type of "bracketing"),[6] whereas in MTT the coordination members are attached one below another.

Finally, it should be mentioned that this paper is written from a humble perspective of a computer scientist interested in Natural Language Processing, with a deep respect to, but with only a limited insight into the two compared approaches. Due to this fact, some parallels outlined in this paper might be inaccurate, which is of course completely our own fault.

Acknowledgements

The presented paper would not be written without the financial support of the project MSM0021620838.

[5] In PDT, also textual coreference is added into the tectogrammatical annotation.

[6] Coordination is treated differently in PDT 2.0: as it is illustrated in Figure 2, there is a special coordination node (functor CONJ), the children of which are the members of coordination.

References

[Apresian et al.2003] Jurij Apresian, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, and Leonid Tsinman. 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT . MTT 2003. First International Conference on Meaning-Text Theory. Paris.

[Babko-Malaya et al.2004] O. Babko-Malaya, M. Palmer, N. Xue, A. Joshi, and S. Kulick. 2004. Proposition Bank II: Delving Deeper. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 17–23, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

[Bolshakov and Gelbukh2000] I. A. Bolshakov and A. F. Gelbukh. 2000. The Meaning-Text Model: Thirty Years After. International Forum on Information and Documentation.

[Cinková2004] Silvie Cinková. 2004. Manuál pro tektogramatickou anotaci angličtiny. Technical report, ÚFAL/CKL MFF UK, Prague.

[Hajič et al.2001] Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. Prague Dependency Treebank 1.0 (Final Production Label). CDROM CAT: LDC2001T10, ISBN 1-58563-212-0.

[Hajič et al.2004] Jan Hajič, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague Arabic Dependency Treebank: Development in Data and Tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117, September 22–23, 2004.

[Kingsbury and Palmer2002] Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of Third International Conference on Language Resources and Evaluation*.

[Mel'čuk and Žolkovskij1984] Igor A. Mel'čuk and Alexander K. Žolkovskij. 1984. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna.

[Mel'čuk1988] Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

[Meyers et al.2004] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An Interim Report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

[Miltsakaki et al.2004] E. Miltsakaki, A. Joshi, R. Prasad, and B. Webber. 2004. Annotating Discourse Connectives and Their Arguments. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

[Nakhimovsky1990] Alexander Nakhimovsky. 1990. Word Meaning and Syntactic Structure: Some Comparative Notes. In *Meaning-Text Theory: Linguistics, Lexicography, and Implications*. University of Ottawa Press.

[Panevová1980] Jarmila Panevová. 1980. *Formy a funkce ve stavbě české věty*. Academia, Praha.

[Sgall et al.1986] Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

[Sgall1967] Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

[Sgall1997] Petr Sgall. 1997. Valency and underlying structure. An Alternative View on Dependency. In L. Wanner, editor, *Recent Trends in Meaning-Text Theory*, pages 149–166. Amsterdam/Philadelphia: Benjamins.

[Straňáková-Lopatková and Žabokrtský2002] Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. 2002. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, volume 3, pages 949–956.

[Tesnière1959] Lucien Tesnière. 1959. *Elments de syntaxe structurale*. Paris.

[Žolkovskij and Mel'čuk1965] Aleksandr Žolkovskij and Igor Mel'čuk. 1965. O vozmoznom metode i instrumentax semanticeskogo sinteza. Naucno-texniceskaja informacija.