Tectogrammatical Representation: Towards a Minimal Transfer In Machine Translation

Jan Hajič
Charles University, Prague, Czech Republic
hajic@ufal.mff.cuni.cz

1. Introduction

The Prague Dependency Treebank (PDT, as described, e.g., in (Hajič, 1998) or more recently in (Hajič, Pajas and Vidová Hladká, 2001)) is a project of linguistic annotation of approx. 1.5 million word corpus of naturally occurring written Czech on three levels ("layers") of complexity and depth: morphological, analytical, and tectogrammatical. The aim of the project is to have a reference corpus annotated by using the accumulated findings of the Prague School as much as possible, while simultaneously showing (by experiments, mainly of statistical nature) that such a framework is not only theoretically interesting but possibly also of practical use.

In this contribution we want to show that the deepest (tectogrammatical) layer of representation of sentence structure we use, which represents "linguistic meaning" as described in (Sgall, Hajičová and Panevová, 1986) and which also records certain aspects of discourse structure, has certain properties that can be effectively used in machine translation[1] for languages of quite different nature at the transfer stage. We believe that such representation not only minimizes the "distance" between languages at this layer, but also delegates individual language phenomena where they belong to - whether it is the analysis, transfer or generation processes, regardless of methods used for performing these steps.

2. The Prague Dependency Treebank

The Prague Dependency Treebank is a manually annotated corpus of Czech. The corpus size is approx. 1.5 million words (tokens). Three main groups ("layers") of annotation are used:

- the morphological layer, where lemmas and tags are being annotated based on their context;
- the analytical layer, which roughly corresponds to the surface syntax of the sentence,
- the tectogrammatical layer, or linguistic meaning of the sentence in its context.

In general, unique annotation for every sentence (and thus within the sentence as well, i.e. for every token) is used on all three layers. Human judgment is required to interpret the text in question; in case of difficult decisions, certain "tie-breaking" rules are in effect (of rather technical nature); no attempt has been made to define what type of disambiguation is "proper" or "improper" at what level.

Technically, the PDT is distributed in text form, with an SGML markup throughout. Tools are provided for viewing, searching and editing the corpus, together with some basic Czech analysis tools (tokenization, morphology, tagging) suitable for various experiments. The data in the PDT are organized in such a way that statistical experiments can be easily compared between various systems - the data have been pre-divided into training and two sets of test data.

In the present section, we describe briefly the Prague Dependency Treebank structure and its history.

2.1. Brief History of the PDT

The Prague Dependency Treebank project has started in 1996 formally as two projects, one for specification of the annotation scheme, and another one for its immediate "validation" (i.e., the actual treebanking) in the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics at Charles University, Prague. The annotation part itself has been carried out in its Linguistic Data Lab. There has been broad cooperation at

* Supported by the Ministry of Education of the ČR Project LN00A0063 and by the NSF Grant 0121285.

1. We suppose the "classic" design of an MT system, namely, Analysis - Transfer - Synthesis (Generation). Although we believe that overall, our representation goes further than many other syntactico-semantic representations of sentence structure, we are far from calling it an interlingua, since it *can* in general have different realization in different languages for the same sentence.

 Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6), pp. 101–110. Universitá di Venezia.

(1)	Od	vlády	čekáme	autonomní	ekologickou	politiku
	od	vláda	čekat	autonomní	ekologická	politika
	RR--2--------	NNFS2-----A--	VB-P---1P-AA-	AAFS4----1A--	AAFS4----1A--	NNFS4-----A--

'From the-government we-are-awaiting an-autonomous environment policy'

Figure 1: Example morphological annotation: form, lemma, tag

the beginning of the project, especially with the Institute of the Czech National Corpus which (in a similar vein to the British National Corpus) has been constituted at the time as the primary site for collection of and public access to large amounts of Czech contemporary texts[2]. A preliminary version of the PDT (called "PDT 0.5") has been released in the summer of 1998, the first version containing the full volume of morphological and analytical annotation has been published by the LDC in the fall of 2001 (Hajič *et al.*, 2001). The funding for the project which currently concentrates on the tectogrammatical layer of annotation as described below is secured through 2004.

2.2. The Morphological Layer

The annotation at the morphological layer is an unstructured classification of the individual tokens (words and punctuation) of the utterance into morphological classes (*morphological tag*s) and *lemma*s. The original word form is preserved, too, of course; in fact, every token has gotten its unique ID within the corpus for obvious reference reasons. Sentence boundaries are preserved and/or corrected if found wrong (as taken from the Czech National Corpus).

There is nothing unexpected at this level of annotation, since it follows closely the design of the Brown Corpus and of the tagged WSJ portion of the Penn Treebank. However, since it is a corpus of Czech, the tagset size used is 4257, with about 1100 different tags actually appearing in the PDT. The data has been double-annotated fully manually, our morphological dictionary of Czech (Hajič, 2001) has been used for generating a possible list of tags for each token from which the annotators selected the correct interpretation.

There are 13 categories used for morphological annotation of Czech: Part of speech, Detailed part of speech, Gender, Number, Case, Possessor's Gender and Number, Person, Tense, Voice, Degree of Comparison, Negation and Variant. In accordance with most annotation projects using rich morphological annotation schemes, so-called positional tag system is used, where each position in the actual tag representation corresponds to one category (see Fig. 1).

2.3. The Analytical Layer

At the analytical layer, two additional attributes are being annotated:

- (surface) sentence structure,
- analytical function.

A single-rooted *dependency tree* is being built for every sentence[3] as a result of the annotation. Every item (token) from the morphological layer becomes (exactly) one node in the tree, and no nodes (except for the single "technical" root of the tree) are added. The *order* of nodes in the *original sentence* is being preserved in an additional attribute, but non-projective constructions are allowed (and handled properly thanks to the original token serial number). Analytical functions, despite being kept at nodes, are in fact names of the dependency relations between a dependent (child) node and its governor (parent) node. As stated above, only one (manually assigned) analytical annotation (dependency tree) is allowed per sentence.

According to the pure dependency tradition, there are no "constituent nodes"[4], as opposed e.g. to the mixed representations in the NEGRA corpus (Skut *et al.*, 1997) which contains the head annotation alongside the constituent structure; we are convinced the constituent nodes are in general not needed for deeper analysis, even though we found experimentally that for parsing, some of the annotation typically found at the constituent level might help

2. The ICNC has now over 0.5 billion words of Czech text available.
3. Sentence-break errors are manually corrected at the analytical layer as well.
4. And no equivalent markup either.

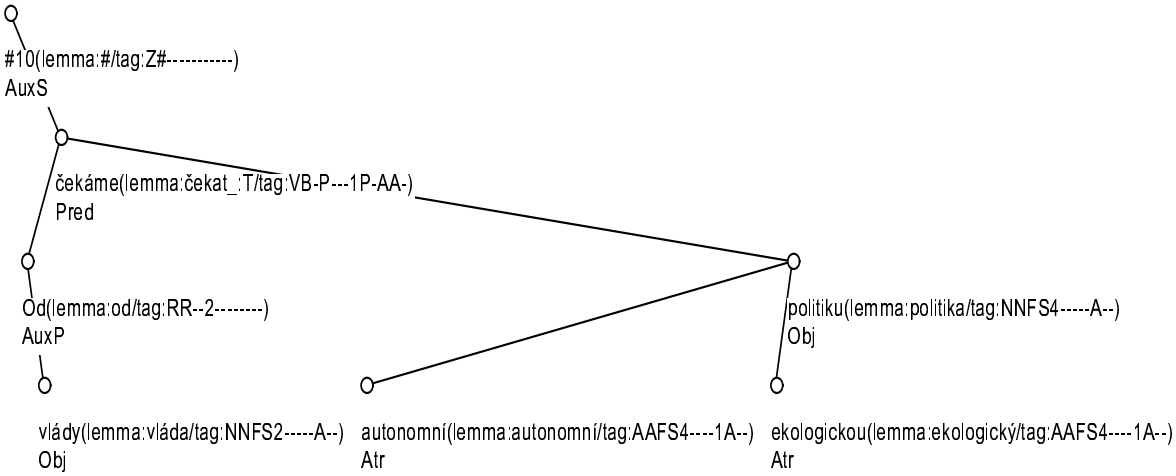

Figure 2: Analytical annotation (sentence from Fig. 1): form, function (+ dependencies, preserved word order)

(such as subordinate clause root markup; for more details, see (Collins *et al.*, 1999)). However, there are still many "technical" dependencies left - we are here at the level of the surface syntax, and there is often no linguistic reason to create a dependency between e.g. an analytical verb form, or a punctuation and everything else, etc.

Coordination and apposition is handled using such "technical" dependencies, too: the conjunction is the head and the members are its "dependent" nodes. Common modifiers of the coordinated structure are also dependents of the coordinating conjunction, but they are not marked as coordinated structure members. This additional "coordinated structure member" markup (`_Co`, `_Ap`) gives an added flexibility for handling such constructions.

Ellipsis is not annotated at this level (no traces, no empty nodes etc.), but a special analytical function (`ExD`) is used at nodes that are lacking their governor, even though they (technically) do have a governor node in the annotation[5].

There are 24 analytical functions used[6], such as `Sb` (Subject), `Obj` (Object, regardless of whether the direct, indirect, etc.), `Adv` (Adverbial, regardless of type), `Pred,Pnom` (Predicate / Nominal part of a predicate for the (verbal) root of a sentence), `Atr` (Attribute in noun phrases), `Atv,AtvV` (Verbal attribute / Complement), `AuxV` (auxiliary verb - similarly for many other auxiliary-type words, such as prepositions (`AuxP`), subordinate conjunctions (`AuxC`), etc.), `Coord,Apos` (coordination/apposition "head"), `Par` (Parenthesis head), etc.

A simple example of the analytical level annotation of the sentence from Fig. 1 is in Fig. 2.

2.4. The Tectogrammatical Layer

The tectogrammatical layer is the most elaborated, complicated but also the most theoretically based layer of syntactico-semantic (or "deep syntactic") representation. The tectogrammatical layer annotation scheme is divided into four sublayers:

- dependencies and functional annotation,
- the topic/focus annotation including reordering according to the deep word order,
- coreference,
- the fully specified tectogrammatical annotation (including the necessary grammatical information).

As an additional data structure we use a syntactic lexicon, mainly capturing the notion of *valency*. The lexicon is not needed for the interpretation of the tectogrammatical representation itself[7], but it is helpful when working on

5. It is the (recursively) closest parent that is physically present in the original sentence.
6. Not counting the additional coordination and special parenthetical markup which effectively triples that number.
7. Nor for further analysis (say, a logical one) based on it, nor (in the other direction) for generation (synthesis) of surface sentences.

the annotation since it defines when a particular node should be created that is missing on the surface. In other words, the notion of (valency-based) ellipsis is defined by the dictionary. But before describing the dictionary, let us describe the first ("core") sublayer of annotation.

Dependencies and Functors

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as "subject" and "object" by notions like "actor", "patient", "addressee" etc. The representation itself still relies upon the language structure itself rather than on world knowledge. The nodes in the tectogrammatical tree are *autosemantic words* only.[8]. Dependencies between nodes represent the relations between the (autosemantic) words in a sentence, for the predicate as well as any other node in the sentence. The dependencies are labeled by *functors* [9], which describe the dependency relations. Every sentence is thus represented as a dependency tree, the nodes of which are autosemantic words, and the (labeled) edges name the dependencies between a dependent and its governor.

Many nodes found at the morphological and analytical layers disappear [10] (such as function words, prepositions, subordinate conjunctions, etc.). The information carried by the deleted nodes is not lost, of course: the relevant attributes of the autosemantic nodes they belong to now contain enough information (at least theoretically) to reconstruct them.

Ellipsis is being resolved at this layer. Insertion of (surface-)deleted nodes is driven by the notion of *valency* (see below the section on Dictionary) and completeness (albeit not in its mathematical sense): if a word is deemed to be used in a context in which some of its valency frames applies, then all the frame's obligatory slots are "filled" (using regular dependency relations between nodes) by either existing nodes or by newly created nodes, and these nodes are annotated accordingly. Textual ellipsis (often found in coordination, direct speech etc.) [11] is resolved by creating a new node and copying all relevant information from its origin, keeping the reference as well.

Every node of the tree is furthermore annotated by such a set of grammatical features that enables to fully capture the meaning of the sentence (and therefore, to recover - at least in theory - the original sentence or a sentence with synonymous linguistic meaning).

The Dictionary (Syntactic, Valency Lexicon)

The tectogrammatical layer dictionary is viewed mainly as a valency dictionary of Czech. By *valency* (as theoretically defined in (Panevová, 1975); for recent account of the computational side and the actual dictionary creation, see (Skoumalová, Straňáková-Lopatková and Žabokrtský, 2001)) we mean the necessity and/or ability of (autosemantic) words to take other words as their dependents, as defined below.

Every dictionary entry is called a *lexia*, which may contain one or more alternative *(valency) frames*. A frame consists of a set of *(valency) slots*. Each slot contains a *function* section (the actual functor, and an indication whether the functor is obligatory [12]), and an associated *form* section. The form section has no direct relation to the tectogrammatical representation, but it is an important link to the analytical layer of annotation: it contains an (underspecified) analytical tree fragment that conforms to the analytical representation of a possible expression of the particular slot. Often, the form section is as simple as a small (analytical) subtree with one (analytical) dependency only, where the dependent node has a particular explicitly specified morphemic case [13]; equally often, it takes the form of a two-edge subtree with two analytical dependencies: one for a preposition (together with its case subcategorization) as the dependent for the surface realization of the root of the lexia itself, and one for the preposition's dependent (which is completely underspecified). However, the form section can be a subtree of any complexity, as it might be the case for phrasal verbs with idiomatic expressions etc.

Moreover, the form section might be different for different expressions (surface realizations) of the lexia itself. For example, if the lexia is a verb and its surface realization is in the passive voice, the form of the (analytical) nodes corresponding to its (tectogrammatical) valency slots will be different than if realized in the active voice.

8. By "autosemantic", as usual, we mean words that have lexical meaning, as opposed to just grammatical function.
9. At two levels of detail; here we ignore so-called *syntactic grammatemes*, which provide the more detailed subclassification.
10. Based on the principle of using only autosemantic words in the representation.
11. Nominal phrases, as used in headings, sports results, artifact names etc. are not considered incomplete sentences, even though they do not contain a predicate.
12. By "obligatory" we mean that this functor (slot) must be present at the tectogrammatical layer of annotation; this has immediate consequences for ellipsis annotation, cf. below.
13. Czech has seven morphemic cases: nominative, genitive, dative, accusative, vocative, locative, and instrumental, usually numbered 1 to 7. In the example in Fig. 1, the case takes the 5^{th} position in the positional representation of the tag.

However, relatively simple rules do exist to "convert" the active forms into the passive ones that work for most verbs; therefore, for such verbs, only the canonical (active) forms [14] are associated with the corresponding valency slots. For irregular passivization problems there is always the possibility to include the two (or more) different realizations explicitly into the dictionary.

A similar mechanism is in place for nominalizations. Verbal nouns typically share the function section of the valency frame with their source verbs, but the form section might be a regular or an irregular transform of the corresponding form section. Again, if the necessary transformation is regular, only the canonical form section need to be present (or even no frame at all, if the verb-to-noun derivation is regular in the function section as well).

Other issues are important in the design of the valency lexicon as well, such as reciprocity, information about verbs of control (Panevová, Řezníčková and Urešová, 2002), etc., but they are outside the scope of this rather brief discussion.

The issue of word sense(s) is not really addressed in the valency dictionary. Two lexias might have exactly the same set of valency frames (as defined above, i.e. including the form section(s) of the slot(s)); in such a case, it is assumed that the two words have different lexical meaning (polysemy) [15]. It is rather practical to leave this possibility in the dictionary (however "dirty" this solution is from the purist syntactic viewpoint), since it allows to link the lexias by a single reference to, e.g. the Wordnet senses (Pala and Ševeček, 1999). The lexical (word sense) disambiguation is, however, being solved outside the tectogrammatical level of annotation, even though eventually we plan to link the two, for obvious reasons. Then it will be possible to relate the lexias for one language to another in their respective (valency) dictionaries (at least for the majority of entires). From the point of view of machine translation, this will serve as an additional source of syntactically-based information of form correspondence between the two languages.

Topic, Focus and Deep Word Order

Topic and focus (Hajičová, Partee and Sgall, 1998) are marked, together with so-called deep word order reflected by the order of nodes in the annotation, is in general different from the surface word order, and all the resulting trees are projective by the definition of deep word order.

By *deep word order* we mean such (partial) ordering of nodes at the tectogrammatical layer that puts the "newest" information to the right, and the "oldest" information to the left, and all the rest inbetween, in the order corresponding to the notion of "communicative dynamism". Such an ordering is fully defined at each single-level subtree of the tectogrammatical tree; i.e., all sister nodes *together with their head* are fully ordered left-to-right. The order is relative to the immediate head only; therefore, there exists such a total ordering of the whole tectogrammatical tree that the tree is projective. We believe that the deep word order is language-universal for every utterance in the same context, unless, roughly speaking, the structural differences are "too big" (or, in other words, the corresponding translation is "too free").

In written Czech, the surface word order roughly corresponds to the deep word order (with the notable systematic exception of adjectival attributes to nouns, and some others), whereas the grammar of English syntax dictates in most cases a fixed order, and therefore the deep word order will be more often different (even though not always; even English has its means to shuffle words around to make the surface word order closer to the deep one, such as extraposition).

Coreference

Grammatical and some textual coreference is resolved and marked. This is subject to future work, despite some ongoing test annotation. Grammatical coreference (such as the antecedent of "which", "whom", etc., control etc.) is simpler and therefore we believe it will be done more easily and sooner that its textual counterpart. (For more on control in PDT, see (Panevová, Řezníčková and Urešová, 2002) in this volume.)

3. Machine Translation and the Tectogrammatical Layer

The usual scenario of machine translation is Analysis - Transfer - Synthesis (Generation). It is commonly accepted wisdom that the deeper the analysis, the smaller the transfer and vice versa. It is equally clear that

14. By "form" we mean the analytical tree fragment as defined above.
15. On the other hand, it is clear that two lexias that do *not* share the same set of frames must have different lexical meaning as well, unless truly synonymous at a higher level of analysis.

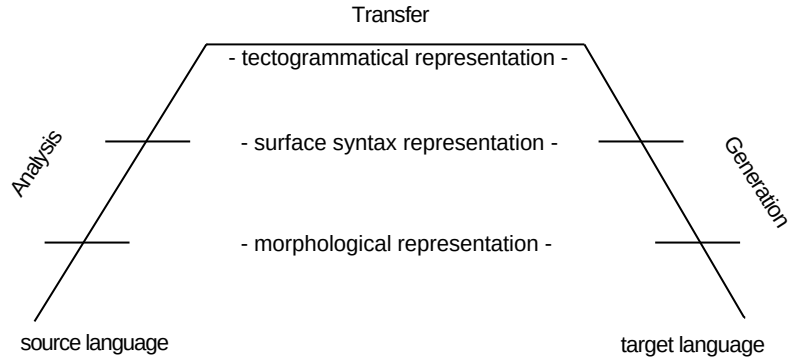

Figure 3: Transfer-based MT scheme with three levels of analysis and generation

the deeper the analysis (and smaller and simpler the transfer), the longer the path from the source to the target languages, and therefore the more errors are likely to creep in. We in principle agree with this, since only careful experiments and variety of evaluations must be run to prove or disprove this. We would like to argue at this point, however, that (even though we have not done such convincing experiments yet), intuitively, there must be an advantage if the transfer end points are defined at a locally clean information saddle point with as least "dirt" from the other language as possible. There has been a number of attempts to use syntactic structure of a sentence to do MT; recently, the most succesful one is statistically based (Yamada and Knight, 2001). We propose here, however, to go to a "deeper" level of analysis.

3.1. The Overall Design

Fig. 3 shows the overall scheme of a transfer-based approach to machine translation. This triangle-based scheme[16] is currently considered the common scheme of all machine translation systems, whether they are of commercial nature (such as (Flanagan and McClure, 2002)) or of research nature ((Brown *et al.*, 1993), (Knight, 1999)) and regardless of their prevailing methodology (with the exception of very few interlingua-based systems (Cavalli-Sforza *et al.*, 2000)).

As Fig. 3 suggests, we propose three essential analysis steps and three generation steps:

- Morphological processing;
- Analytical (surface syntax) processing;
- Tectogrammatical processing (underlying syntax);

and, of course, transfer at the top of the processing "triangle"[17].

An output from one step is the input to the following step; thus we have here four representations of the data along the "up-leading" as well as the "down-leading" paths (from bottom to top):

- The *surface form* of the text (i.e., the actual input and output of the whole system).
- Unstructured *morphological representation* (cf. Sect. 2.2), i.e., an ordered list of lemmas and morphological tags. The order corresponds to the original word order of the sentence.
- Structured *analytical representation* (cf. Sect. 2.3), in the form of a dependency tree that contains all tokens from the morphological layer. Let's summarize that every token is annotated by the lemma and tag coming from the morphological layer, and by a pointer to its governing node and an analytical function naming the dependency relation. The left-to-right order of the nodes of the tree is still coming from the surface sentence word order, therefore causing non-projective trees at times.

16. We should rather call it a "trapezoid" scheme, since the top is always cut off in it.
17. Word sense disambiguation (WSD) is not considered a separate step in this scenario, but of course it is taken care of at the tectogrammatical representation level, unless it is already solved while parsing to the tectogrammatical level (based on different valency frames of the words in question).

- Structured *tectogrammatical representation* (for more details and the four sublayers of annotation, cf. Sect. 2.4) which does *not* contain the word form, lemma, morphological tag, analytical function, nor the surface dependency links. Instead, the tectogrammatical dependency, lexia [18] and the functor is used as the basic information here, supplemented by grammatemes that contain information about number, tense, degree of comparison only where it cannot be recovered from the lexia and function itself. In the full tectogrammatical representation, coreference and deep word order together with topic/focus is annotated as well.

Let us now illustrate how the correspondence among quite distant languages (English, Czech and Arabic) becomes more and more apparent (and straightforward) as we move up the translation "triangle". We will use the sentence *The only remaining baker bakes the most famous rolls north of Long River*, which translates to Czech as *Jediný zbývající pekař peče nejznámější rohlíky na sever od Dlouhé řeky* and to Arabic as (transcribed) *'al-xabbaaz 'al-'axiir 'al-baaqii yaśnacu 'ashhar 'al-kruasaanaat ilaa shimaal min Long River*.

3.2. Surface Form and Morphological Layer Correspondence

Even though the example sentence is quite straightforward to translate (certainly more easily than many sentences in the WSJ), it is clear that there are several unpleasant (non-)correspondences at the surface form, and similarly at the morphological level: articles have no correspondence in the Czech sentence, whereas in the Arabic counterpart, articles are in fact part of the Arabic words. Similarly, the superlative is expressed in Czech by circumfixing, whereas in English it is represented by several words and in Arabic there is a specific single word (*'ashhar*). The Arabic word order is different, too: the word for "baker" (*'al-xabbaaz*) precedes its attributes in the Arabic translation, but follows them in both Czech and English. Therefore methods based on very shallow analysis (i.e., morphological at most) will have trouble (at least) with different word counts, different word order, and, as usual, lexical choice (cf. further below).

3.3. Analytical Layer Correspondence

Fig. 4 shows the corresponding trees. The correspondence of the dependencies is more visible, but since the number of nodes is the same as on the morphological layer, the problems mentioned above did not disappear; on the contrary, the surface structure of the Arabic superlative construction (*'ashhar 'al-kruasaanaat*) even reverses the associated dependency relation (compared to both Czech and English, cf. *the most famous rolls*). Since the original word order is preserved in the analytical dependency tree, the shape of the tree does not correspond even for simple nominal phrases [19]. Overall, even though many dependencies do correspond to each other, there are still many dependencies that either do not correspond to anything in the other language, or are reversed.

3.4. Tectogrammatical Correspondence

Even though there is some similarity between languages at the surface dependency syntax level, the tectogrammatical structure displays often striking similarity, both in the structure and in the functor correspondence (Fig. 5), even though we say again that it is not meant to be an [artificial] interlingua [20].

Analytical Verb Forms

Verbs tend to use various auxiliaries to express person, tense, sometimes number and other morphological properties. We believe, however, that once the person, tense, number etc. is determined (disambiguated), then there is no need to have separate nodes for each of the auxiliaries. The auxiliaries are completely determined by the language in question; therefore, we must be able to handle the insertion of appropriate auxiliaries at the generation stage, which is by its nature already monolingual. In the context of machine translation, this is especially useful: we have to take care of the main (autosemantic) verb only during the translation proper (the transfer phase), but not of the (presence or absence) of auxiliary source words. For example, the type of auxiliary in German perfect tense (*sein*/*haben*) is grammatically (or, lexically) based and has nothing to do with the other language, since that language might use quite different auxiliaries (or none at all, if it uses inflection to express the perfect tense).

18. Recall that "lexia" is the lexical unit reference at the tectogrammatical level, and thus it plays here a role similar to the "lemma" at the morphological and analytical layers.
19. Although the direction of dependencies does (*remaining* depends on *baker*, similarly in Arabic *'al-baaqii* depends on *'al-xabbaaz* and in Czech *zbývající* depends on *pekař*).
20. For example, compare the difference in the structure for "I like swimming" in English and "Ich schwimme gern" in German.

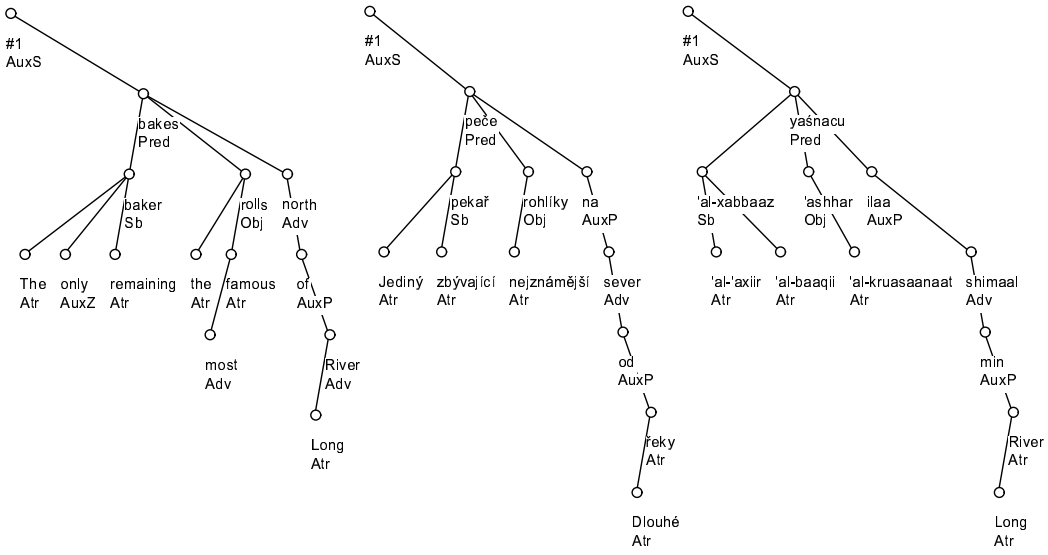

Figure 4: Analytical layer correspondence: *The only remaining baker bakes the most famous rolls north of Long River* in English, Czech and Arabic

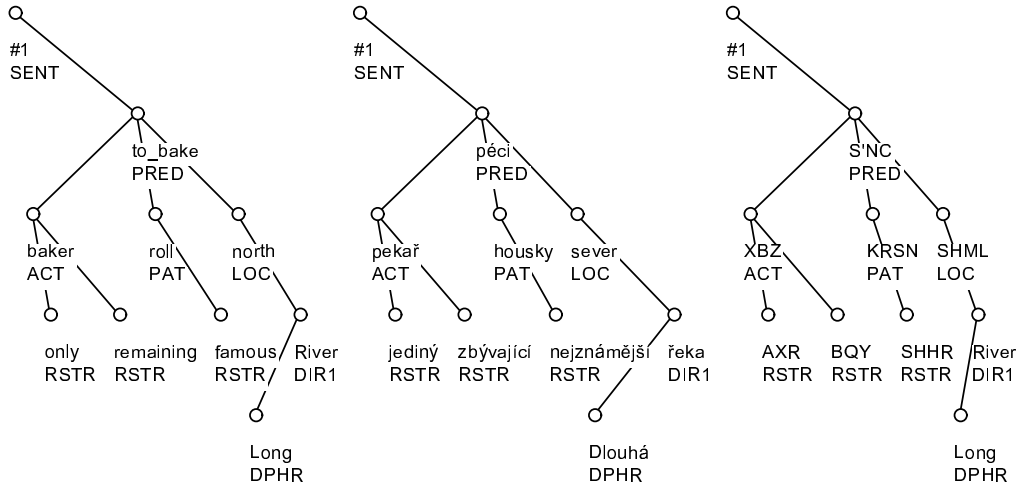

Figure 5: Tectogrammatical layer correspondence: *The only remaining baker bakes the most famous rolls north of Long River*, again in English, Czech and Arabic; for transparency reasons, only the lexia and functor are shown, and not grammatemes such as number.

Articles

Articles pose a difficult problem for translation into a language that has articles (such as English) from a language that does not (such as Czech). The choice of an article is hardly conditioned by words of the other language; rather, it is either determined grammatically, referentially, or by the topic/focus distinction described earlier. We thus believe that articles, as another class of non-autosemantic words, has no place at the tectogrammatical layer of representation, since the topic/focus and deep word order should be sufficient (together with the grammar rules of the target language) to insert the right articles at the right places. For example, the need to use "the" in front of every superlative is purely English-grammar-related and certainly does not stem from the language being translated from or to; the choice of "a" vs. "the" for a general word such as "keyboard" will be determined by the topic/focus annotation: if the word "keyboard" is in the topic, the definite article ("the") typically has to be used, otherwise "a" should be used instead.

Choice of Prepositions (and Morphemic Case)

Prepositions usually do exist across languages, even though they are not always used as separate words (cf. Hungarian and other agglutinative languages), and often a "default" translation can be found for every preposition. However, from the experience with inflective languages such as Czech, we consider prepositions and morphemic cases to be at the same "level" - if not just a form variant - expressing a particular tectogrammatical functor [21]. Therefore, when translating into English, we have to select prepositions, when translating e.g. into Czech we have to decide the case or preposition [22].

Even then, the relation between functors and prepositions/cases is not always straightforward, for at least two reasons:

- The choice of preposition is driven by usage in the target language (e.g., it depends on the noun used with the preposition or on some similar factor);
- The choice of preposition/case is driven by the governing word *and* by the functor of the dependent word (i.e. the one that has to get the preposition/case).

In both cases, the source language sentence representation does not help much. In the first case, we simply have to have a language model or similar knowledge of the target language [23] that simulates usage. In the second case, a valency dictionary of the target language (as defined in Sect. 2.4) comes in handy: once we are able to determine the correct target word (more precisely, the lexia as the translation of the source lexia), a valency dictionary entry gives matching functors and with each of them, its surface expression (by means of an underspecified analytical-level annotated subtree, mostly either just a case or a preposition with its own subcategorization for a morphemic case).

For words not having valency, their dependent nodes (as well as dependent nodes of all words with non-valency modifier functors) acquire their preposition or case as the default value for each functor.

Word Order

Word order differs across languages, of course, sometimes wildly. English has its word order mostly grammatically given (meaning that the grammar dictates that sentences should in the SVO order, that the rules for systemic ordering must be followed, etc.); some exceptions in the grammar do allow for some word shuffling, such as extraposition. However, Czech word order is discourse-driven (and thus not so "free" as often mislabeled). The correct solution, in our opinion, to the word order problem is thus not to deal with it at the transfer level, but at the analysis level (determining the deep word order), and at the generation stage (using the determined deep word order to perhaps generate an extraposition, and using the grammar rules [24] of the target language to determine the correct word order).

21. Some regular "transformations" notwithstanding, such as in passivization, where the surface syntax expression also plays a role.
22. Prepositions have subcategorization for case, so for subcategorization-ambiguous prepositions the correct subcat frame must be selected together with the preposition.
23. A good language model (as used in automatic speech recognition systems) can actually help in many cases of target-language-related conditioning.
24. By grammar rules we mean here any kind of "rules"; it is expected that these rules will be learned within a statistical modeling framework.

Agreement (in the Generation Stage)

Grammatical agreement is again determined by the rules of the target language, and not by the translation itself. Its importance in English is low, obviously, but it is crucial for other languages. E.g., in Czech, every adjective has to agree in gender, number and case with its head noun. We propose to deal with this problem at the analytical level, once the analytical tree is built (which includes solving the word order issue, of course); it is not related to the tectogrammatical level in fact. Thus, for example, only the number is needed to be preserved (translated) at the tectogrammatical level[25], its dependent adjectives will be the populated by the correct morphemic values once also the case is determined by the rules described above, and once the gender of the noun is determined from the lexicon. The formation of the surface text is then easy through any morphological generator of the target language, since the word order has been defined in the preceding stages.

4. Conclusion

We have described the basic ideas and annotation scheme for the Prague Dependency Treebank, a reference corpus with three-level linguistic annotation for morphology, surface syntax, and so-called tectogrammatical layer representation. We have then argued that the tectogrammatical layer is suitable not only for various linguistic experiments, but also for practical use, specifically for machine translation systems, since it generalizes (and disambiguates) in such a way that it achieves - to a certain extent limited by "language meaning" - independence of both the source and target languages. We believe that our representation has the potential to improve the overall translation quality, and that the additional burden of deeper analysis will not outweight its benefits.

References

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2).

Cavalli-Sforza, Violetta, Krzystof Czuba, Teruko Mitamura and Eric Nyberg. 2000. Challenges in Adapting an Interlingua for Bidirectional English-Italian Machine Translation. In *AMTA 2000*.

Collins, Michael, Jan Hajič, Eric Brill, Lance Ramshaw and Christopher Tillmann. 1999. A Statistical Parser for Czech. In *37th Meeting of the Association of Computational Linguistics*, pages 505–512, University of Maryland, College Park, MD, June 22nd–25th.

Flanagan, Mary and Steve McClure. 2002. SYSTRAN and the Reinvention of MT. In electronic form only, at SYSTRAN's web pages (`http://www.systransoft.com/IC/26459.html`).

Hajič, Jan. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Festschrift for Jarmila Panevová*. Karolinum, Charles University Press, Prague, pages 106–132.

Hajič, Jan. 2001. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Prague, Czech Republic: Faculty of Math. and Physics, Charles University. hab. thesis.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall and Barbora Vidová Hladká. 2001. The Prague Dependency Treebank. CD-ROM Catalog #LDC2001T10. ISBN 1-58563-212-0.

Hajič, Jan, Petr Pajas and Barbora Vidová Hladká. 2001. The Prague Dependency Treebank: Annotation Structure and Support. *IRCS Workshop on Linguistic Databases*, pages 105–114.

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

Knight, Kevin et al. 1999. EGYPT: a statistical machine translation toolkit. Summer Workshop'99, Johns Hopkins University, Baltimore, MD, `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit`.

Pala, Karel and Pavel Ševeček. 1999. Final Report EuroWordNet-1,2, project LE4-8328. Technical report, EU Commission, Amsterdam, Sept. 1999. On the project CD-ROM.

Panevová, Jarmila. 1975. On Verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics (PBML)*, 22 (Part I), 23 (Part II):3–40,17–52.

Panevová, Jarmila, Veronika Řezníčková and Zdeňka Urešová. 2002. The Theory of Control Applied in Tagging of the Prague Dependency Treebank. In Robert Frank, editor, *TAG+6 Workshop (this volume)*, Venice, May 20-23, 2002. Univ. of Pennsylvania.

Sgall, Petr, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of a Sentence in its Semantic and Pragmatic Aspects*. Prague - Amsterdam: Academia - North–Holland.

Skoumalová, Hana, Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. 2001. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In *Text, Speech and Dialogue, Lecture Notes on Computer Science LNCS 2166*, pages 142–147, Železná Ruda, Czech Rep., Sept. 2001. Springer-Verlag.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit. 1997. An Annotataion Scheme for Free Word Order Languages. In *the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, Washington, D.C., USA, March 1997. Association of Computational Linguistics.

Yamada, Kenji and Kevin Knight. 2001. A Syntax-Based Statistical Translation Model. In *39th Meeting of the Association of Computational Linguistics*, Toulouse, France, July 2001.

25. One of the so-called grammatemes - invisible in our examples above - is devoted to this.