Prague Arabic Dependency Treebank: Research Directions

Otakar Smrž
Džám-e Džam Language Institute
Czech Republic
otakar.smrz@seznam.cz

Abstract

In this contribution, we briefly describe the methods and contents of the Prague Arabic Dependency Treebank. We then outline some important directions of research and application development that can hopefully come to the foreground in the field and be pursued by the open scientific community as this novel linguistic resource is completed, or when similar Arabic computational linguistics projects are made publicly available.

1 Introduction

Prague Arabic Dependency Treebank (PADT) is a collection of linguistically annotated texts from various Arabic newspapers and news agencies (Parker et al., 2009). The supplied linguistic annotation makes explicit the morphological properties of words and the syntactic structures of sentences. In a subset of the texts, it also formally represents their deep linguistic meaning.

The research context of the PADT project is most accurately described in (Smrž et al., 2008). The finalization of the second release of PADT is still in progress, even though it is expected in a couple of months time. The new release will include not only valuable linguistic annotations, but even a powerful suite of tools for browsing and processing the data. PADT is closely connected with excellent open-source projects, such as the TrEd/PML-TQ (Pajas and Štěpánek, 2008; Pajas and Štěpánek, 2009) annotation environment and complex treebank data management system, or the ElixirFM (Smrž and Bielický, 2010; Bielický and Smrž, 2009) computational morphology and lexicon of Modern Written Arabic.

The outcomes of the PADT project can find application in various areas of natural language processing, linguistics, and education, as already confirmed by the interest in the first release of PADT (Hajič et al., 2004). The initial version of PADT covered over one hundred thousand words of text, whereas the new release will exceed one million words annotated with morphology and syntax.

PADT is maintained by the Institute of Formal and Applied Linguistics, Charles University in Prague. The website http://ufal.mff.cuni.cz/padt/online/ offers further information on current developments of the project, as well as contact details.

2 Annotation Levels

Prague Arabic Dependency Treebank comprises refined linguistic annotations whose style is influenced by the Functional Generative Description theory (Sgall et al., 1986; Hajičová and Sgall, 2003) and by the Prague Dependency Treebank project (Hajič et al., 2006). The multi-level description scheme discerns functional morphology, analytical dependency syntax, and tectogrammatical representation of linguistic meaning.

Morphological annotations identify the textual forms of a discourse lexically and recognize their grammatical properties. The analytical syntactic processing describes the superficial dependency structures in the discourse, while tectogrammatics reveals the underlying dependency structures and restores linguistically relevant semantic information. Figure 1 parses this example sentence:

> في ملف الأدب طرحت المجلة قضية اللغة
> العربية والأخطار التي تهددها.
>
> In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it.

2.1 Functional Morphology

Morphological and phonological processing of the Arabic language is considered challenging not only for the templatic nature of the structure of words, but also for the properties of the Arabic script into which words are normally encoded.

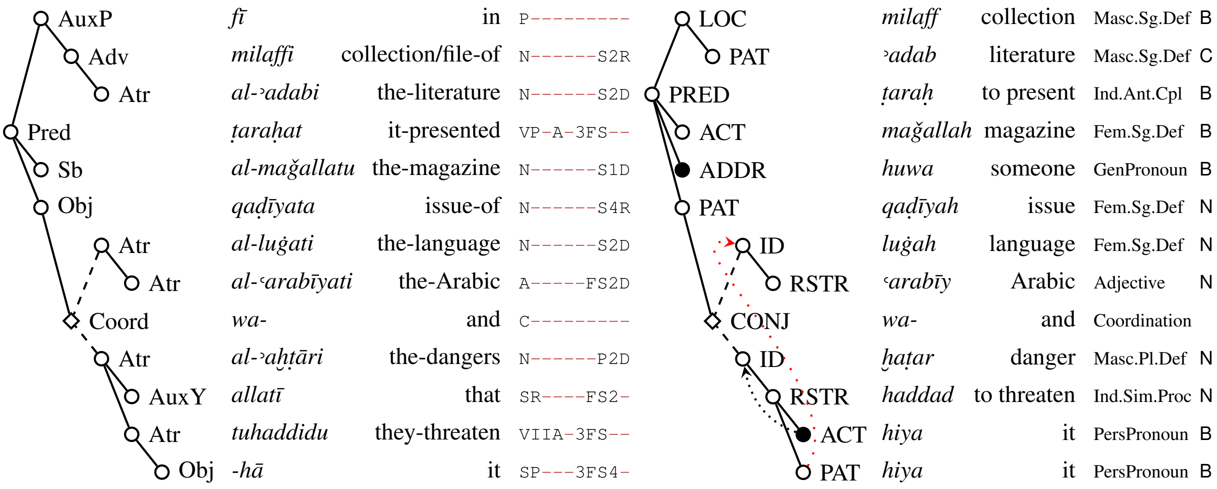
AuxP fī in P---------
Adv milaffi collection/file-of N------S2R
Atr al-ʾadabi the-literature N------S2D
Pred ṭaraḥat it-presented VP-A-3FS--
Sb al-maǧallatu the-magazine N------S1D
Obj qaḍīyata issue-of N------S4R
Atr al-luġati the-language N------S2D
Atr al-ʿarabīyati the-Arabic A-----FS2D
Coord wa- and C---------
Atr al-ʾaḫṭāri the-dangers N------P2D
AuxY allatī that SR----FS2-
Atr tuhaddidu they-threaten VIIA-3FS--
Obj -hā it SP---3FS4-
LOC milaff collection Masc.Sg.Def B
PAT ʾadab literature Masc.Sg.Def C
PRED ṭaraḥ to present Ind.Ant.Cpl B
ACT maǧallah magazine Fem.Sg.Def B
ADDR huwa someone GenPronoun B
PAT qaḍīyah issue Fem.Sg.Def N
ID luġah language Fem.Sg.Def N
RSTR ʿarabīy Arabic Adjective N
CONJ wa- and Coordination
ID ḫaṭar danger Masc.Pl.Def N
RSTR haddad to threaten Ind.Sim.Proc N
ACT hiya it PersPronoun B
PAT hiya it PersPronoun B

Figure 1: *Left:* Example of analytical annotation. Orthographic words are tokenized into lexical words, and their inflectional morphosyntactic properties are encoded using positional tags. Members of coordination are depicted with dashed edges. *Right:* Example of tectogrammatical annotation with resolved coreference (extra arcs) and indicated values of contextual boundness. Lexemes are identified by citation forms, and selected grammatemes are shown in place of morphosyntactic features.

Many attempted approaches to Arabic morphology do not succeed to provide exact and clear, yet generally fitting and extensible models of word formation, since they cannot disentangle the whole morphological process properly into its independent, simpler, well-defined components.

The functional view of language pursued in PADT requires a morphological model capable of more appropriate and deeper generalizations than what the popular Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) or other systems convey. Earlier functional approximations used in PADT, which were derived by imperfectly tweaking the output of (Buckwalter, 2002), are now replaced by the solutions of the innovative ElixirFM system (Smrž, 2007; Smrž and Bielický, 2010).

ElixirFM is suited for both morphological analysis and generation, and can be used as an advanced multi-purpose morphological model. In the interactive mode, one can invoke various utility functions for lookup in the lexicon, inflection and derivation of lexemes, resolution of strings, exporting and pretty-printing of the information, etc., as well as explore the definitions of the underlying linguistic rules and data being involved. The ElixirFM source code and the lexicon itself are highly reusable by both computers and humans.

Word forms are explicit in their morphological structure. They are specified via the underlying template of morphs and the inherited root. Merging the template with the root produces the form in the ArabTEX notation (Lagally, 2004), from which the orthographic string or its phonetic version can be generated, cf. Figure 2.

ElixirFM carefully designs the morphophonemic patterns of the templates, as well as the phonological rules hidden in the `>|` or `|<<` operators. This greatly simplifies the morphological rules proper, both inflectional and derivational. Inspired by functional programming in Haskell (Forsberg and Ranta, 2004), ElixirFM implements many generalizations of classical grammars (Fischer, 2002; Ryding, 2005), and suggest even some new abstractions (Smrž, 2007).

2.2 Surface Syntax

Annotations on the analytical level are represented by dependency trees. Their nodes map, one to one, to the tokens resulting from the morphological analysis and tokenization, and their roots group the nodes according to the division into sentences or paragraphs. Edges in the trees show there is a syntactic relation between the governor and its dependent, or rather, the whole subtree under and including the dependent. The nature of the government is expressed by the analytical functions of the nodes being linked, e.g. Subject, Object, Attribute.

2.3 Deep Syntax

Tectogrammatics, the underlying syntax reflecting the linguistic meaning of an utterance, is the highest level of annotation in the family of Prague

```
|> "d r y" <| [                                                       d r y دري
FaCY                                 `verb`  [ "know", "notice" ]             faʿā
      `imperf`  FCI                                                            fʿī
      `masdar`  FiCAL |< aT,                                              fiʿāl-ah
FACY                                 `verb`  [ "flatter", "deceive" ],        fāʿā
HaFCY                                `verb`  [ "inform", "let know" ],        ʾafʿā
TaFACY                               `verb`  [ "hide", "conceal" ],          tafāʿā
lA >| "'a" >>| FCI |<< "Iy" |< aT    `noun`  [ "agnosticism" ],       lā-ʾa-fʿī-īy-ah
lA >| "'a" >>| FCI |<< "Iy"          `adj`   [ "agnostic" ],              lā-ʾa-fʿī-īy
FiCAL |< aT                          `noun`  [ "knowledge", "knowing" ],   fiʿāl-ah
MuFACY |< aT                         `noun`  [ "flattery" ]                mufāʿā-ah
      `plural`  MuFACY |< At,                                              mufāʿā-āt
HaFCY                                `adj`   [ "more knowledgeable" ],        ʾafʿā
FACI                                 `adj`   [ "aware", "knowing" ] ]          fāʿī
```

know, notice	(*dirāyah* دراية)	I (*i*)	*darā* درى	agnostic		*lā-ʾadrīy* لاأدري
flatter, deceive		III	*dārā* دارى	knowledge, knowing		*dirāyah* دراية
inform, let know		IV	*ʾadrā* أدرى	flattery	(*mudārayāt* مداريات)	*mudārāh* مداراة
hide, conceal		VI	*tadārā* تدارى	more knowledgeable		*ʾadrā* أدرى
agnosticism			*lā-ʾadrīyah* لاأدرية	aware, knowing		*dārin* دار

Figure 2: Excerpt from the ElixirFM lexicon and a layout generated from it. The source code of entries nested under the *d r y* دري root is shown in the typewriter font. Note the custom notation specifying the underlying morphological structure of words and the economy yet informativeness of the declarations.

Dependency Treebanks (Hajič et al., 2006). It captures dependency and valency (Žabokrtský, 2005; Bielický and Smrž, 2008) with respect to the deep linguistic relations of discourse participants. In its generality, the description also includes topic–focus articulation, coreference resolution, and other non-dependency relations.

The topology of a tectogrammatical representation of a sentence is similar to that of the analytical level. In contrast to it, nodes in the tree may be deleted, inserted, and even reorganized. We speak of a transfer of structures from analytical to tectogrammatical, which can be partly automated.

Tectogrammatical nodes appear as lexical entries rather than inflected forms. Grammatemes, the deep grammatical parameters, abstract away from the morphological and analytical features of an utterance. Functors, the deep roles that the participants assume, include Actor, Patient, Addressee, Origin, Effect, various types of local and temporal modifications, Extent, Manner, Cause, Identity, Restriction, coordination types, and many more (Mikulová and others, 2006).

Figure 1 compares the analytical and tectogrammatical representations of the example sentence. The black inserted nodes are recovered from the discourse, since they are obligatory arguments of the valency frames of the two verbal predicates. Values of contextual boundness, a feature from which the topic–focus dichotomy is inferred, are also indicated (Hajičová and Sgall, 2003).

3 Research Directions

With the availability of large annotated treebank data, new topics for computational linguistics research are opening. PADT is certainly not the only Arabic treebank around, cf. (Maamouri and Bies, 2004; Habash et al., 2009; Dukes and Buckwalter, 2010), not to mention several other derived or converted Arabic treebanks. However, the PADT annotation levels are unique for the kind and extent of information they contain.

Let us outline a couple of thoughts about the new research topics and applications for which the PADT data and related tools can be very use-

```
|> "` q d" <| [
    FaCaL      `verb`  [ "convene", "conclude" ]  `imperf`  FCiL,
    FaCCaL     `verb`  [ "complicate" ],
    TaFaCCaL   `verb`  [ "be complicated" ],
    TaFACaL    `verb`  [ "contract", "convene" ],
    InFaCaL    `verb`  [ "be held", "be gathered", "be convened" ],
    IFtaCaL    `verb`  [ "believe" ] ]
```

		ʿ *q d* عقد
ʿaqad	I (*i*)	عَقَد
ʿaqqad	II	عَقَّد
taʿaqqad	V	تَعَقَّد
taʿāqad	VI	تَعَاقَد
inʿaqad	VII	إِنعَقَد
iʿtaqad	VIII	إِعتَقَد

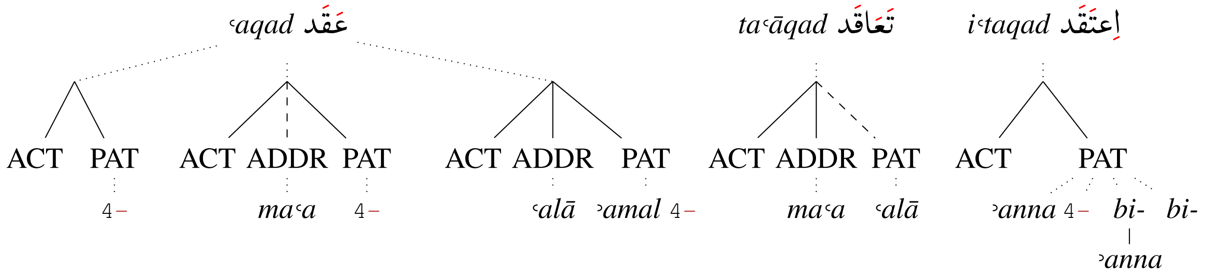

Figure 3: *Top left:* Verb entries of the ElixirFM lexicon nested under the ʿ *q d* عقد root. *Top right:* Possible layout of these entries including the explicit derivational class, showing that various pieces of information can be inferred directly from this lexicon's representation. *Bottom:* Valency frame treelets and the constraints on the surface realization of the functors, organized into trees. Optional slots are marked with dashed edges. Multiple options with frames or constraints are rendered as dotted links.

ful. We do not discuss here the development of statistical parsers and taggers, since such systems have already been tackled, though not solved completely, cf. (Smrž et al., 2008).

3.1 Language Generation

Describing linguistic structures in an appropriate formal system can serve not only for representing the meaning of utterances. It also allows generating the natural language, as well as transforming it and translating it. In syntax-driven machine translation, sentences in the source language are parsed into their grammatical representations before the translation of the syntactic structure is performed. Then, the structures are "spelled out" or linearized into the target language.

The problem of language generation can be seen as an inverse to linguistic parsing—language generation is a function from some structured representation of linguistic meaning into a linear sequence of graphemes or phonemes used in a natural language to express the meaning. The particular instance of the generation problem and its complexity therefore depends both on the character of information supplied as input, and on the requirements on the form of the output.

There are several generic approaches that implement the overall translation framework as well as parsing of e.g. English sentences. The task then would be to implement the translation of the "universal" syntactic structures into Arabic syntactic structures and then to produce the Arabic word forms required by the morphosyntactic parameters implied by the structures.

Examples of the linguistic structures assumed as the parameters for generation into Arabic can be obtained from the family of Prague Dependency Treebanks, cf. (Hajič et al., 2006; Smrž et al., 2008), can be annotated after their guidelines, or can be transformed from other sources. Interestingly, the same kind of structures can be produced automatically during a transfer-based machine translation processing chain (Žabokrtský et al., 2008; Žabokrtský et al., 2010).

The annotated data of PADT as well as the ElixirFM system itself can be used directly for the morphological generation. What would remain as the essential problem is the pruning of multiple possible translations, both for every word and for the whole sentences. One then needs to score possible solutions and optimize for the best one. However, this is a language-independent problem for which there exist published methods and implementations. Why not try that?

Arabic language generation has been addressed previously by (Soudi, 2004; Dada, 2007; Habash et al., 2007), among others. Arabic morphological generation, in particular, is treated quite often, cf. (Beesley, 1996; Cavalli-Sforza et al., 2000; Habash, 2004; Habash et al., 2005; Altantawy et al., 2010). None of the works, though, combines real-world language resource like PADT with a working, open-source software implementation modeling both inflectional and derivational morphological processes, providing lexicon lookup based on concrete as well as abstract search criteria, and offering systematic linguistic resolution of Arabic either in form of the running text, or in phonetic transcription or other notations, like ElixirFM does.

The ElixirFM lexicon is now enriched with valency frames of selected verbal lexemes, cf. Figure 3 (Bielický and Smrž, 2009). Not only does a valency frame—represented as a tree structure of alternative dependency subtrees—encode the valency properties of a lexeme in terms of the functors, i.e. the underlying syntactic roles. A valency frame also provides possible morphemic representations (e.g. prepositions, conjunctions) and morphosyntactic features (grammatical case or state) of the particular argument or complementation when realized on the surface.

Valency frames stored in the ElixirFM lexicon can positively contribute to effective generation of sentences, since they combine the underlying tectogrammatical information with the requirements for their surface representation.

3.2 Lexical Semantics

This project would be concerned with taking the open-source lexical resources that are available for Arabic, like (Buckwalter, 2002; Smrž and Bielický, 2010; Dukes and Buckwalter, 2010), as well as in combination with multiple other languages, possibly extracted from parallel data.

The point of the project would be to build lexical networks, using links to the other language resources as well as links within words of the language itself, discovered through unsupervised methods of machine learning (Church and Hanks, 1990; Brown et al., 1992; Evert, 2005). One would need to use information theory and graph theory for this. The outcome would be new improved Arabic lexicons that would feature multi-word lexical items, synonyms, antonyms, hyponyms and hypernyms, etc.

4 Conclusion

We have presented two important resources that we have co-authored, namely the Prague Arabic Dependency Treebank and the ElixirFM morphological system for Arabic. Both resources are highly reusable, not only for language analysis and parsing, but also for language generation with all its applications. We have proposed and discussed novel research directions that seem to be ready to be pursued by the Arabic computational linguistics community, in part due to these two resources.

References

Mohamed Altantawy, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Kenneth R. Beesley. 1996. Arabic Finite-State Morphological Analysis and Generation. In *Proceedings of the 16th conference on Computational linguistics*, pages 89–94, Morristown, NJ, USA. Association for Computational Linguistics.

Viktor Bielický and Otakar Smrž. 2008. Building the Valency Lexicon of Arabic Verbs. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Viktor Bielický and Otakar Smrž. 2009. Enhancing the ElixirFM Lexicon with Verbal Valency Frames. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. LDC2002L49, ISBN 1-58563-257-0.

Violetta Cavalli-Sforza, Abdelhadi Soudi, and Teruko Mitamura. 2000. Arabic Morphology Generation Using a Concatenative Strategy. In *Proceedings of NAACL 2000*, pages 86–93, Seattle.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Ali Dada. 2007. Implementation of the Arabic Numerals and their Syntax in GF. In *ACL 2007 Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 9–16, Prague.

Kais Dukes and Tim Buckwalter. 2010. A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings of the 7th international Conference on Informatics and Systems (INFOS 2010)*.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institute for Natural Language Processing, University of Stuttgart.

Wolfdietrich Fischer. 2002. *A Grammar of Classical Arabic*. Yale University Press.

Markus Forsberg and Aarne Ranta. 2004. Functional Morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference on Functional Programming, ICFP 2004*, pages 213–223. ACM Press.

Nizar Habash, Owen Rambow, and George Kiraz. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24.

Nizar Habash, Ryan Gabbard, Owen Rambow, Seth Kulick, and Mitch Marcus. 2007. Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features. In *Proc. of EMNLP-CoNLL 2007*, pages 1084–1092.

Nizar Habash, Reem Faraj, and Ryan Roth. 2009. Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*.

Nizar Habash. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In *JEP-TALN 2004, Session Traitement Automatique de l'Arabe*, Fes, Morocco.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kráčmar, and Kamila Hassanová. 2004. Prague Arabic Dependency Treebank 1.0. LDC2004T23, ISBN 1-58563-319-4.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN 1-58563-370-4.

Eva Hajičová and Petr Sgall. 2003. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I, pages 570–592. Walter de Gruyter.

Klaus Lagally. 2004. ArabTEX: Typesetting Arabic and Hebrew, User Manual Version 4.00. Technical Report 2004/03, Fakultät Informatik, Universität Stuttgart, March 11.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *COLING 2004 Computational Approaches to Arabic Script-based Languages*, pages 2–9, Geneva.

Marie Mikulová et al. 2006. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Technical Report 30, UFAL MFF UK, Charles University in Prague.

Petr Pajas and Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 673–680.

Petr Pajas and Jan Štěpánek. 2009. System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2009. Arabic Gigaword Fourth Edition. LDC2009T30, 1-58563-532-4.

Karin C. Ryding. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel & Academia.

Otakar Smrž and Viktor Bielický. 2010. ElixirFM. Functional Arabic Morphology, http://sourceforge.net/projects/elixir-fm/.

Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague Arabic Dependency Treebank: A Word on the Million Words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco.

Otakar Smrž. 2007. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague.

Abdelhadi Soudi. 2004. Challenges in the Generation of Arabic from Interlingua. In *JEP-TALN 2004, Session Traitement Automatique de l'Arabe*, Fes, Morocco.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *StatMT '08: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170.

Zdeněk Žabokrtský, Martin Popel, and David Mareček. 2010. Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–206.

Zdeněk Žabokrtský. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University in Prague.