

Information Structure with the Prague Arabic Dependency Treebank

Otakar Smrž Petr Zemánek Jakub Kráčmar Viktor Bielický

Institute of Formal and Applied Linguistics
& Institute of Comparative Linguistics
Charles University in Prague

Conference on Communication and Information Structure
in Spoken Arabic

The issue of information structure in language has been studied extensively both in the **Prague School of Linguistics** and in the **Functional Generative Description** [Sgall et al., 1986, Hajičová and Sgall, 2003].

This theory of **representation of linguistic meaning** is the framework for a family of **multi-level annotation** projects, esp. Prague Dependency Treebank for Czech [Hajič et al., 2001, 2006] and **Prague Arabic Dependency Treebank** [Hajič et al., 2004, Smrž et al., 2006].

The issue of information structure in language has been studied extensively both in the **Prague School of Linguistics** and in the **Functional Generative Description** [Sgall et al., 1986, Hajičová and Sgall, 2003].

This theory of **representation of linguistic meaning** is the framework for a family of **multi-level annotation** projects, esp. Prague Dependency Treebank for Czech [Hajič et al., 2001, 2006] and **Prague Arabic Dependency Treebank** [Hajič et al., 2004, Smrž et al., 2006].

Outline

- 1 Introduction
- 2 Definitions
 - dependency / information structure
 - communicative dynamism / valency
 - topic–focus / contextual boundness
- 3 Examples
 - systemic ordering
 - topicalizers
 - rhematizers
- 4 Annotation
- 5 Discussion

Outline

- 1 Introduction
- 2 Definitions
 - dependency / information structure
 - communicative dynamism / valency
 - topic–focus / contextual boundness
- 3 Examples
 - systemic ordering
 - topicalizers
 - rhematizers
- 4 Annotation
- 5 Discussion

Outline

- 1 Introduction
- 2 Definitions
 - dependency / information structure
 - communicative dynamism / valency
 - topic–focus / contextual boundness
- 3 Examples
 - systemic ordering
 - topicalizers
 - rhematizers
- 4 Annotation
- 5 Discussion

Outline

- 1 Introduction
- 2 Definitions
 - dependency / information structure
 - communicative dynamism / valency
 - topic–focus / contextual boundness
- 3 Examples
 - systemic ordering
 - topicalizers
 - rhematizers
- 4 Annotation
- 5 Discussion

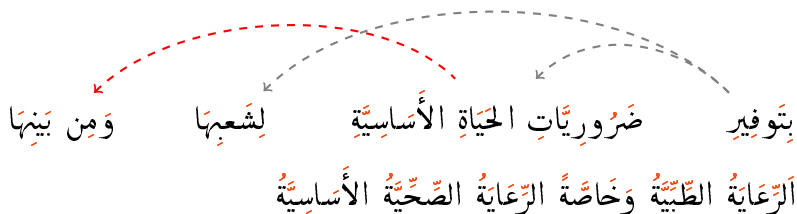
Outline

- 1 Introduction
- 2 Definitions
 - dependency / information structure
 - communicative dynamism / valency
 - topic–focus / contextual boundness
- 3 Examples
 - systemic ordering
 - topicalizers
 - rhematizers
- 4 Annotation
- 5 Discussion

Dependency Description

Representation of **structure** in language built on **dependency** exhibits

- Immediate Dominance relation
- Linear Precedence relation



*bi-tawfīri ḍarūrīyāti 'l-ḥayāti 'l-ʾasāsīyati li-šaʿbihā wa-min baynihā
ar-riʿāyatu 'ṭ-ṭibbīyatu wa-ḥāṣṣatan ar-riʿāyatu 'ṣ-ṣiḥḥīyatu 'l-ʾasāsīyatu*

Dependency Description

Representation of **structure** in language built on **dependency** exhibits

- Immediate Dominance relation
- Linear Precedence relation

tirǧa^c tgul li ʔumm l-bnayya

ʔumm l-bnayya tirǧa^c tgul li

[Brustad, 2000, page 336]

تَرْجِعْ تَكُولِ لِي أُمُّ الْبَنِيَّةِ

أُمُّ الْبَنِيَّةِ تَرْجِعْ تَكُولِ لِي

يَا أُمَّ أَحْمَدَ أَنْتِ تَعْرِفِينَهُ؟

Dependency Description

Representation of **structure** in language built on **dependency** exhibits

- Immediate Dominance relation
- Linear Precedence relation

tirǧa^c tgul li ʔumm l-bnayya

ʔumm l-bnayya tirǧa^c tgul li

[Brustad, 2000, page 336]

tirǧa^c ʔumm l-bnayya tgul li

تَرْجِعْ تَكُولِ لِي أُمُّ الْبَنِيَّةِ

أُمُّ الْبَنِيَّةِ تَرْجِعْ تَكُولِ لِي

يَا أُمَّ أَحْمَدَ أَنْتِ تَعْرِفِيْنَهُ؟

تَرْجِعْ أُمُّ الْبَنِيَّةِ تَكُولِ لِي

Dependency Description

Representation of **structure** in language built on **dependency** exhibits

- Immediate Dominance relation
- Linear Precedence relation

Contrast dependency with constituency, i.e. **phrase-structure** syntax, where linear order and derivation are not independent, and thus less expressive (context-free vs. context-sensitive).

Dependency Description

Representation of **structure** in language built on **dependency** exhibits

- Immediate Dominance relation
- Linear Precedence relation

Contrast dependency with constituency, i.e. **phrase-structure** syntax, where linear order and derivation are not independent, and thus less expressive (context-free vs. context-sensitive).

Distinction of different points of view and **levels of abstraction**

- Surface/Analytical syntax level
- Deep/Tectogrammatical syntax level

Description includes recovery of **tree topology** and **syntactic functions**.

Information Structure

Information structure — the question of “the given” and “the new” in an utterance and how it is expressed — is considered to contribute to the linguistic meaning, and its annotation is thus in our interest.

Information Structure

Information structure — the question of “the given” and “the new” in an utterance and how it is expressed — is considered to contribute to the linguistic meaning, and its annotation is thus in our interest.

How is **communicative dynamism** expressed and delivered/packaged?

- word order variation with respect to **systemic ordering**
- intonation centers, sentence stress
- extra constructs in syntax or morphology

an tanḥafīḍa qīmatu ... أَنْ تَنْخَفِضَ قِيَمَةَ الصَّادِرَاتِ إِلَى الْوَلَايَاتِ الْمُتَّحِدَةِ
fī 'n-niṣfi ... فِي النِّصْفِ الثَّانِي مِنْ السَّنَةِ الْجَارِيَةِ
ilā 400 ... min 593 ... إِلَى ٤٠٠ مِلْيُونِ دُولَارٍ مِنْ ٥٩٣ مِلْيُونِ دُولَارٍ حَالِيًا
bi-sababi ... بِسَبَبِ تَدَاعِيَاتِ الْأَحْدَاثِ

Topic / Focus

Dichotomy of **aboutness** inspiring to many similar concepts in modern linguistics [cf. Kruijff-Korbayová and Steedman, 2003, for overview].

Topic (theme) part of sentence structure **linking** the **content** of the utterance with the **context** of the discourse

Focus (rheme, comment) the part **providing** or **modifying** some **information** about the topic

Topic / Focus

Dichotomy of **aboutness** inspiring to many similar concepts in modern linguistics [cf. Kruijff-Korbayová and Steedman, 2003, for overview].

Topic (theme) part of sentence structure **linking** the **content** of the utterance with the **context** of the discourse

Focus (rheme, comment) the part **providing** or **modifying** some **information** about the topic

Communicative dynamism as measure of this linguistic property

- Topic proper is the least dynamic part
- Focus proper is the most dynamic part

Contextual Boundness

The **elementary distinction**, from which topic–focus dichotomy is derived, is **contextual boundness**.

Context-Bound lexical reference to an already **explicitly mentioned** entity, or to an entity **implicitly evoked** in the context of the discourse

Non-Bound lexical item that is not contextually bound, i.e. **not available** in the interlocutor's mind as reference

Contextual Boundness

The **elementary distinction**, from which topic–focus dichotomy is derived, is **contextual boundness**.

Context-Bound lexical reference to an already **explicitly mentioned** entity, or to an entity **implicitly evoked** in the context of the discourse

Non-Bound lexical item that is not contextually bound, i.e. **not available** in the interlocutor's mind as reference

Using **question test** to **identify** the context-bound and non-bound items.

Contextual Boundness

The **elementary distinction**, from which topic–focus dichotomy is derived, is **contextual boundness**.

Context-Bound lexical reference to an already **explicitly mentioned** entity, or to an entity **implicitly evoked** in the context of the discourse

Non-Bound lexical item that is not contextually bound, i.e. **not available** in the interlocutor's mind as reference

Using **question test** to **identify** the context-bound and non-bound items.

The relation of **definiteness** and boundness is **not trivial** [Kruijff-Korbyová, 1998, Brustad, 2000]. Boundness is **not equated** to the **given/new** concept.

عَلَى عَوْدَةٍ الشَّيْخِ إِلَى الْكُوَيْتِ مِنْ لَنْدُنْ

‘alā ‘awdati ‘š-šayhi ‘ilā ‘l-kuwayti min landn

يَتِمُّ الْقَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu ‘l-qabḍu ‘alayhim min ḥīnin ‘ilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ الْعَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-ma‘nan ‘āḥara ‘inna dawranā ‘l-‘asāsīya huwa ‘l-‘arḍu wa-laysa ‘l-intiqādu

فِيمَا يَلِي الْحَقَائِقُ وَالْأَرْقَامُ لِوَارِدَاتِ الصَّيْنِ الرَّئِيسِيَّةِ

fīmā yalī ‘l-ḥaqā‘iqu wa-‘l-‘arqāmu li-wāridāti ‘š-šīni ‘r-ra‘īsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الْحِزْبِ الْحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī ‘s-sābiqi ‘alā ‘l-ḥizbi ‘l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ الْعُنْفِ

alladī yas‘ā faqaṭ ‘ilā waqfi ‘l-‘unfi

عَلَى عَوْدَةٍ الشَّيْخِ إِلَى الكُوَيْتِ مِنْ لَنْدُن

‘alā ‘awdati ‘š-šayhi ‘ilā ‘l-kuwayti min landn

يَتِمُّ القَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu ‘l-qabḍu ‘alayhim min ḥīnin ‘ilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ العَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-ma‘nan āḥara ‘inna dawranā ‘l-‘asāsīya huwa ‘l-‘arḍu wa-laysa ‘l-intiqādu

فِيمَا يَلِي الحَقَائِقُ وَالْأَرْقَامُ لِوَارِدَاتِ الصِّينِ الرَّئِيسِيَّةِ

fīmā yalī ‘l-ḥaqā‘iqu wa-‘l-‘arqāmu li-wāridāti ‘š-šīni ‘r-ra‘īsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الحِزْبِ الحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī ‘s-sābiqi ‘alā ‘l-ḥizbi ‘l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ العُنْفِ

alladī yas‘ā faqaṭ ‘ilā waqfi ‘l-‘unfi

عَلَى عَوْدَةِ الشَّيْخِ إِلَى الْكُوَيْتِ مِنْ لَنْدُنْ

ʿalā ʿawdati 'š-šayhi ʾilā 'l-kuwayti min landn

يَتِمُّ الْقَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu 'l-qabḍu ʿalayhim min ḥīnin ʾilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ الْعَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-maʿnan ʾāḥara ʾinna dawranā 'l-ʾasāsīya huwa 'l-ʿarḍu wa-laysa 'l-intiqādu

فِيمَا يَلِي الْحَقَائِقُ وَالْأَرْقَامُ لِوَارِدَاتِ الصَّيْنِ الرَّئِيسِيَّةِ

fīmā yalī 'l-ḥaqāʾiqu wa-'l-arqāmu li-wāridāti 'š-šīni 'r-raʾīsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الْحِزْبِ الْحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī 's-sābiqi ʿalā 'l-ḥizbi 'l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ الْعُنْفِ

alladī yasʿā faqaṭ ʾilā waqfi 'l-ʿunfi

عَلَى عَوْدَةٍ الشَّيْخِ إِلَى الْكُوَيْتِ مِنْ لَنْدُنْ

‘alā ‘awdati ‘š-šayhi ‘ilā ‘l-kuwayti min landn

يَتِمُّ الْقَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu ‘l-qabḍu ‘alayhim min ḥīnin ‘ilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ الْعَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-ma‘nan ‘āḥara ‘inna dawranā ‘l-‘asāsīya huwa ‘l-‘arḍu wa-laysa ‘l-intiqādu

فِيمَا يَلِي الْحَقَائِقَ وَالْأَرْقَامَ لِوَارِدَاتِ الصِّينِ الرَّئِيسِيَّةِ

fīmā yalī ‘l-ḥaqā‘iqu wa-‘l-‘arqāmu li-wāridāti ‘š-šīni ‘r-ra‘īsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الْحِزْبِ الْحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī ‘s-sābiqi ‘alā ‘l-ḥizbi ‘l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ الْعُنْفِ

alladī yas‘ā faqaṭ ‘ilā waqfi ‘l-‘unfi

عَلَى عَوْدَةِ الشَّيْخِ إِلَى الْكُوَيْتِ مِنْ لَنْدُنْ

‘alā ‘awdati ‘š-šayḥi ‘ilā ‘l-kuwayti min landn

يَتِمُّ الْقَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu ‘l-qabḍu ‘alayhim min ḥīnin ‘ilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ الْعَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-ma‘nan ‘āḥara ‘inna dawranā ‘l-‘asāsīya huwa ‘l-‘arḍu wa-laysa ‘l-intiqādu

فِيمَا يَلِي الْحَقَائِقُ وَالْأَرْقَامُ لِوَارِدَاتِ الصِّينِ الرَّئِيسِيَّةِ

fīmā yalī ‘l-ḥaqā‘iqu wa-‘l-‘arqāmu li-wāridāti ‘s-šīni ‘r-ra‘īsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الْحِزْبِ الْحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī ‘s-sābiqi ‘alā ‘l-ḥizbi ‘l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ الْعُنْفِ

alladī yas‘ā faqaṭ ‘ilā waqfi ‘l-‘unfi

عَلَى عَوْدَةٍ الشَّيْخِ إِلَى الكُوَيْتِ مِنْ لَنْدُنْ

‘alā ‘awdati ‘š-šayḥi ‘ilā ‘l-kuwayti min landn

يَتِمُّ القَبْضُ عَلَيْهِمْ مِنْ حِينٍ إِلَى حِينٍ

yatimmu ‘l-qabḍu ‘alayhim min ḥīnin ‘ilā ḥīnin

بِمَعْنَى آخَرَ إِنَّ دَوْرَنَا الْأَسَاسِيَّ هُوَ العَرْضُ وَلَيْسَ الْاِنْتِقَادُ

bi-ma‘nan ‘āḥara ‘inna dawranā ‘l-‘asāsīya huwa ‘l-‘arḍu wa-laysa ‘l-intiqādu

فِيمَا يَلِي الحَقَائِقُ وَالْأَرْقَامُ لِوَارِدَاتِ الصِّينِ الرَّئِيسِيَّةِ

fīmā yalī ‘l-ḥaqā‘iqu wa-‘l-‘arqāmu li-wāridāti ‘s-šīni ‘r-ra‘īsīyati

الَّتِي كَانَتْ تَقْتَصِرُ فِي السَّابِقِ عَلَى الحِزْبِ الحَاكِمِ فَقَطْ

allatī kānat taqtaṣiru fī ‘s-sābiqi ‘alā ‘l-ḥizbi ‘l-ḥākimi faqaṭ

الَّذِي يَسْعَى فَقَطْ إِلَى وَقْفِ العُنْفِ

alladī yas‘ā faqaṭ ‘ilā waqfi ‘l-‘unfi

The Annotation

Relevance Issue

Topic–focus articulation is relevant even **semantically** — it affects the truth value of a proposition:

- presupposition, allegation, meaning proper
- scope of quantifiers, scope of negation

Relevance Issue

Topic–focus articulation is relevant even **semantically** — it affects the truth value of a proposition:

- presupposition, allegation, meaning proper
- scope of quantifiers, scope of negation

The **applicability** of the general approach to **written** as well as **spoken** Arabic becomes the main point of our account. In FGD, the description of information structure is related also the notions of **intonation center** and stress, **contrast**, subjective **word order**, potential ellipsis, **prosody**, . . .

Relevance Issue

Topic-focus articulation is relevant even **semantically** — it affects the truth value of a proposition:

- presupposition, allegation, meaning proper
- scope of quantifiers, scope of negation

The **applicability** of the general approach to **written** as well as **spoken** Arabic becomes the main point of our account. In FGD, the description of information structure is related also the notions of **intonation center** and stress, **contrast**, subjective **word order**, potential ellipsis, **prosody**, . . .

Annotated **corpora** for written and spoken Arabic are becoming available for quantitative **evaluation** of linguistic **theories**, large-scale analysis of **linguistic** material, **computational** processing and **modeling**.

Conclusion / Prospects

In PADT, which now consists of the **morphological** and the **analytical** levels of description of Arabic, the annotation of **information structure** and **tectogrammatics** is being established.

In our contribution, we have tried to overview the **theoretical concepts** we work with, and present our **formal treatment** of a number of corpus-based instances of linguistic phenomena that have a principal impact on the structure of information in Arabic.

Rich **linguistic literature** and interesting **computational systems** are available [cf. e.g. Hajičová et al., 1995, Kruijff-Korbayová, 1998, Hajičová and Sgall, 2004, Debusmann et al., 2005, Mikulová et al., 2006].

Conclusion / Prospects

In PADT, which now consists of the **morphological** and the **analytical** levels of description of Arabic, the annotation of **information structure** and **tectogrammatics** is being established.

In our contribution, we have tried to overview the **theoretical concepts** we work with, and present our **formal treatment** of a number of corpus-based instances of linguistic phenomena that have a principal impact on the structure of information in Arabic.

Rich **linguistic literature** and interesting **computational systems** are available [cf. e.g. Hajičová et al., 1995, Kruijff-Korbayová, 1998, Hajičová and Sgall, 2004, Debusmann et al., 2005, Mikulová et al., 2006].

Conclusion / Prospects

In PADT, which now consists of the **morphological** and the **analytical** levels of description of Arabic, the annotation of **information structure** and **tectogrammatics** is being established.

In our contribution, we have tried to overview the **theoretical concepts** we work with, and present our **formal treatment** of a number of corpus-based instances of linguistic phenomena that have a principal impact on the structure of information in Arabic.

Rich **linguistic literature** and interesting **computational systems** are available [cf. e.g. Hajičová et al., 1995, Kruijff-Korbayová, 1998, Hajičová and Sgall, 2004, Debusmann et al., 2005, Mikulová et al., 2006].

References I

- Kristen E. Brustad. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press, 2000.
- Ralph Debusmann, Oana Postolache, and Maarika Traat. A Modular Account of Information Structure in Extensible Dependency Grammar. In *Proceedings of the CICLING 2005 Conference*, 2005.
- Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. Prague Dependency Treebank 1.0. LDC catalog number LDC2001T10, ISBN 1-58563-212-0, 2001.
- Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. Prague Dependency Treebank 2.0. LDC catalog number LDC2006T??, ISBN 1-58563-???-?, 2006.
- Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnidauf, Emanuel Beška, Jakub Kráčmar, and Kamila Hassanová. Prague Arabic Dependency Treebank 1.0. LDC catalog number LDC2004T23, ISBN 1-58563-319-4, 2004.
- Eva Hajičová and Petr Sgall. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I, pages 570–592. Walter de Gruyter, 2003.

References II

- Eva Hajičová and Petr Sgall. Degrees of Contrast and the Topic–Focus Articulation. volume 1, pages 1–13. Walter de Gruyter, Berlin, 2004.
- Eva Hajičová, Petr Sgall, and Hana Skoumalová. An Automatic Procedure for Topic–Focus Identification. *Computational Linguistics*, 21(1):81–94, 1995.
- Ivana Kruijff-Korbayová. *The Dynamic Potential of Topic and Focus: A Praguian Approach to Discourse Representation Theory*. PhD thesis, Charles University in Prague, 1998.
- Ivana Kruijff-Korbayová and Mark Steedman. Discourse and Information Structure. *Journal of Logic, Language and Information*, 12(3), 2003.
- Marie Mikulová et al. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Technical report, Charles University in Prague, 2006.
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel & Academia, 1986.
- Otakar Smrž, Petr Pajas, Zdeněk Žabokrtský, Jan Hajič, Jiří Mírovský, and Petr Němec. Learning to Use the Prague Arabic Dependency Treebank. In *Perspectives on Arabic Linguistics*, volume XIX. John Benjamins, 2006.