Multilingual Pragmaticon: Database of Discourse Formulae

Anton Buzanov, Polina Bychkova, Arina Molchanova, Anna Postnikova, Daria Ryzhova

HSE University, Moscow

Outline

- Introduction
- Typological Pragmaticon
- Conclusion

Introduction

Multilingual lexical resources

- (Multi)WordNet
- (Multi)FrameNet
- BabelNet
- MetaNet
- DatSemShift

+ Constructicons!

Discourse formulae (Rakhilina et al. 2021)

- multi-word, (mostly) idiomatic units
- used in a dialogue as a reaction to the speech act of an interlocutor
- express refusal, acceptance, negation, confirmation, disbelief etc.

— I'm making waffles. Want some? — I'm good.

— May I put on some music? — Be my guest.

— We've made it to the finals! — No way!

Russian Pragmaticon (Yaskevitch et al. 2021)

A database of Russian DF designed as a tool for language learners

Available at pragmaticon.ruscorpora.ru

Contains over 600 Russian DF from a list, compiled by manual mark-up of drama and fiction, followed by automatic extraction (Gerasimenko et al. 2020)

Can be used as a starting point for typological comparison

Gathering typological data

- Starting with Slavic languages: supposedly more similar structures

- Finding semantic (pragmatic) counterparts

  - searching parallel corpora

  - comparing contextual distributions with questionnaires (Bychkova 2020, on English data Koziuk, Badryzlova 2021)

- Finding formal counterparts by working through the list with a native speaker

Cross-linguistic variation

- Target pragmatic meanings
  - What meanings are consistently expressed with DF in all the languages?
  - What meanings are unusual for DF?
  - Which pragmatic meanings are contiguous? (polysemy in DF)

- Source meanings
  - Common structures with common meanings
  - "Unique" structures
  - Common structures with different meanings

Typological Pragmaticon

Database structure

- language,
- inner structure,
- glosses,
- lemmas,
- pragmatic function,
- additional semantics,
- speech acts,

- dialog structure,
- intonation,
- syntax,
- source construction,
- source construction syntax
- source construction intonation

Pragmatic functions

- **agreement**
  - Slovene
    sliši se dobro 'sounds good'
- **assessment**
  - Slovene
    brez veze 'not relevant'
- **confirmation**
  - Serbian
    tako-tako 'this is so'
- **indifference**
  - Slovene
    briga me 'not bother me'

- **negation**
  - Russian
    da čto tam 'pointless to discuss'
- **prohibition**
  - Russian
    i ne dumaj 'don't even think'
- **refusal**
  - Serbian
    Bog zna 'lit.: God knows'
- **surprise**
  - Serbian
    Bog s tobom   'holy cow!'
                      lit.: 'God with you'

Additional semantics

- More nuanced semantic characteristics, such as negative or positive assessment, disbelief, doubt or confidence

- Can be combined

a to ty ne znaeš

Pragmatics: **REFUSAL**

Additional semantics: cause: unwillingness | obvious | refusal to answer | verbal

Intonation: polar question

RU

See more

Speech acts and dialog structure

- bipartite

- <polar question>
- a gospod' ego znaet

ptcl God him knows

[**REFUSAL**]

- tripartite

- <question>
- <request to specify the reason>
- da tak

ptcl so

[**REFUSAL**]

Inner structure

Provides a loose description of the literal meaning of a formula. This parameter is two-leveled: the main field corresponds to a more general classification, while the second field highlights additional distinctions.

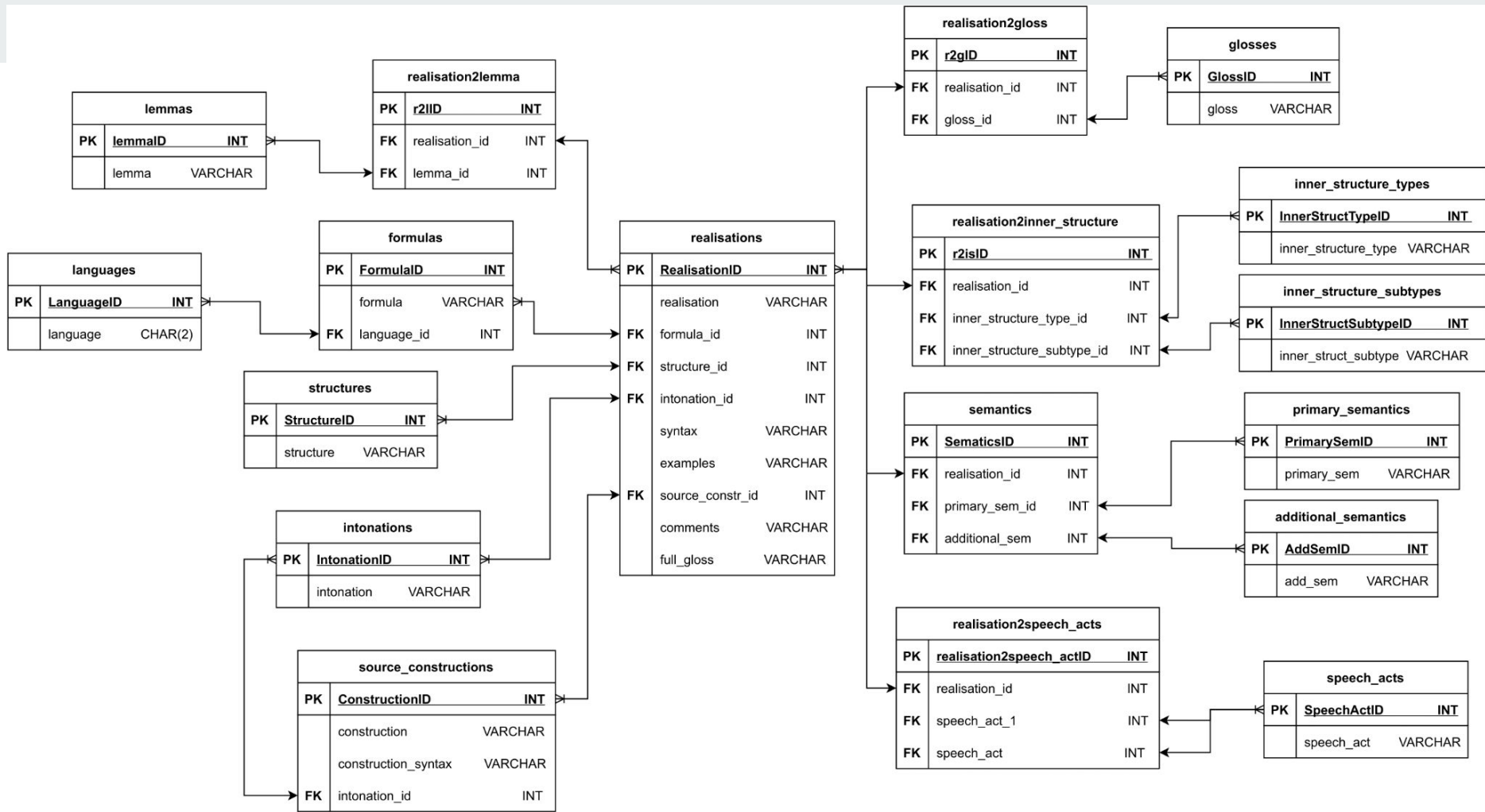**ne možet byt'** — epistemic modality:impossibility

Pragmatics: **NEGATION**

Additional semantics: non-expert

Intonation: wh-question

RU

See more

lemmas

PK	lemmaID	INT
	lemma	VARCHAR

realisation2lemma

PK	r2lID	INT
FK	realisation_id	INT
FK	lemma_id	INT

realisation2gloss

PK	r2gID	INT
FK	realisation_id	INT
FK	gloss_id	INT

glosses

PK	GlossID	INT
	gloss	VARCHAR

languages

PK	LanguageID	INT
	language	CHAR(2)

formulas

PK	FormulaID	INT
	formula	VARCHAR
FK	language_id	INT

structures

PK	StructureID	INT
	structure	VARCHAR

realisations

PK	RealisationID	INT
	realisation	VARCHAR
FK	formula_id	INT
FK	structure_id	INT
FK	intonation_id	INT
	syntax	VARCHAR
	examples	VARCHAR
FK	source_constr_id	INT
	comments	VARCHAR
	full_gloss	VARCHAR

realisation2inner_structure

PK	r2isID	INT
FK	realisation_id	INT
FK	inner_structure_type_id	INT
FK	inner_structure_subtype_id	INT

inner_structure_types

PK	InnerStructTypeID	INT
	inner_structure_type	VARCHAR

inner_structure_subtypes

PK	InnerStructSubtypeID	INT
	inner_struct_subtype	VARCHAR

semantics

PK	SematicsID	INT
FK	realisation_id	INT
FK	primary_sem_id	INT
FK	additional_sem	INT

primary_semantics

PK	PrimarySemID	INT
	primary_sem	VARCHAR

additional_semantics

PK	AddSemID	INT
	add_sem	VARCHAR

intonations

PK	IntonationID	INT
	intonation	VARCHAR

source_constructions

PK	ConstructionID	INT
	construction	VARCHAR
	construction_syntax	VARCHAR
FK	intonation_id	INT

realisation2speech_acts

PK	realisation2speech_actID	INT
FK	realisation_id	INT
FK	speech_act_1	INT
FK	speech_act	INT

speech_acts

PK	SpeechActID	INT
	speech_act	VARCHAR

Example queries

http://pragmaticon.ru

Conclusion

Possible applications

- Studies of pragmaticalization – a process of emergence of pragmatic markers which express the speaker's attitude towards something, including the interlocutor(s) and their speech acts (Diewald 2011)
- Classification of pragmatic meanings built on a typological basis
- Catalog of translational equivalents of the DF in different contexts
- Template for multilingual databases of routines (*God bless you!*) and interjections (*Oh my God!*)

Future plans

Extending the language sample

...so, we invite you to contribute to our database!

References

Aijmer K. (1997). I think – an English modal particle. Modality in Germanic languages: Historical and comparative perspectives, 1–47.

Bast, R., et al. (2021). The Russian Constructicon. An electronic database of the Russian grammatical constructions. Available at https://constructicon.github.io/russian/.

Bybee, J. (2006). From Usage to Grammar: The Mind's Response to Repetition. Language, 82(4), 711–733.

Diewald, G. (2011). Pragmaticalization (defined) as grammaticalization of discourse functions. Linguistics 49 (2011), Nr. 2, 49(2), 365-390.

Gerasimenko, E., et al. (2019). Defining discourse formulae: computational approach. EPiC Series in Language and Linguistics, 4, 61-69.

Haspelmath, M. (1989). From Purposive To Infinitive—A Universal Path Of Grammaticization. Folia linguistica historica, 23, 287–310.

Haspelmath, M. (2004). On directionality in language change with particular reference to grammaticalization.

Hopper & Traugott (2003). Grammaticalization. Cambridge University Press.

Kuteva, T., et al. (2019). World lexicon of grammaticalization. Cambridge University Press.

Narrog, H., & Heine, B. (Eds.). (2011). The Oxford handbook of grammaticalization. Oxford University Press.

Noël, D. (2007). Diachronic construction grammar and grammaticalization theory. In Functions of Language Vol. 14, Issue 2, pp. 177–202.

Onodera N. O. The grammaticalization of discourse markers. The Oxford handbook of grammaticalization, 2011.

Rakhilina, E., et al. (2021) Speech acts as a linguistic category: The case of discourse formulae. Voprosy jazykoznanij, 2, 7-27.

Traugott & Heine (Eds.). (1991). Approaches to Grammaticalization. John Benjamins Publishing.

Yaskevitch, A., et al. (2021). The Russian Pragmaticon. An electronic database of the Russian pragmatic constructions. Available at https://pragmaticon.ruscorpora.ru/

Grammaticalization

"'[G]rammaticalization is [. . .] that subset of linguistic changes through which a lexical item in certain uses becomes a grammatical item, or through which a grammatical item becomes more grammatical". (Hopper & Traugott 2003: 2)

The process of grammaticalization is well studied and widely discussed in literature (Heine & Traugott 1991, Haspelmath 1989, 2004, Narrog & Heine 2011).

"World lexicon of grammaticalization" (Kuteva et al 2019) lists the inventory of lexical items that develop into grammatical markers.

Pragmaticalization (Aijmer 1997, Onodera 2011)

A process of emergence of pragmatic markers which express the speaker's attitude towards something, including the interlocutor(s) and their speech acts (cf. refusal).

This process remains understudied (see, however, Diewald 2011): it is yet to be discovered, which meanings should be considered pragmatic and what the sources for pragmatic units are.

The process should be studied on typological grounds, aiming at "world lexicon of **pragmaticalization**".

Which kind of data could be taken as a basis?

Some observations

DF with verbs of speech

The inner structure tags highlight repetitive patterns:

Russian *čto ty* + V-2p.PRES [verb of speech]

Serbian *šta* + V-2p.PRES [verb of speech]

Slovene *kaj* + V-2p.PRES [verb of speech]

DF appeared to be frozen (partially due to the most frequent ones that are), however bigger data shows variation.

Why so many alternatives that are close both semantically and formally?

Multiple formulae or one schematic construction?

A possible explanation: these are not multiple formulaic phrases but one construction with a slot.

In Russian, there is a standard construction *Čto Pron VP!*

However:

- in DF we encounter only closely synonymous verbs.
  Constructions are normally more productive, cf. *Čto Pron VP!*

  *Čto ty **begaeš'**! Čto vy tut **sidite!** Čto **oni** šumjat! Čto ty **delaeš'**!*

- this group of DF developed further pragmatic meanings:
  Čto Pron VP! — negative assessment / prohibition of the action
  Čto ty govoriš! — negative assessment / prohibition of the speech act
  + surprise, + negation

Initially, there was a construction with a slot, but only a finite number of its realizations developed into DF.

Another explanation

The pairs <DF structure> — <DF meaning> fit into larger, more basic schemas of pragmatic shifts:
<source **meaning**> — <target meaning>.

Unlike grammatical markers that emerge from lexical items, DF emerge from sentences.

There are usually multiple means (compositional and idiomatic) to put the same proposition in a sentence:
cf. *Ne govori erundy* VS. *Ne meli čepuxi* VS. *Čto za erundu ty neseš′* VS. *Ladno erundu-to nesti*

Within one schema of the shift, the **phrasing** of the source meaning can vary, regardless of whether on a level of a lexical variable, or as an entirely diferent construction.

Schemas

A global strategy:

<SOURCE> discontent with interlocutor's speech act => <TARGET> negation

Accommodates multiple types of formally different sources, more or less mutually replaceable:

- verbal prohibition (RU: *perestan'*, *ne govori erundy*, SR: *ne lupaj gluposti*, *nemoj da lupaš*)
- direct assessment of the speech act (RU: *kakaja čepuxa*, SR: *ma to su gluposti*)
- rhetorical question about the speech act (RU: *čto ty govoriš'*; SR: *ma šta pričaš / lupaš*)
- rhetorical question about the interlocutor (RU: *ty s uma sošel*? SR: *jesi li normalan*? SL: *a si nor*?), etc.

Ultimately, these are indirect speech acts: one message conveyed by means of another.

The more conflict the situation, the bigger variety of indirect means, cf. requests:

V! Would you V? / Could you V? / Will you V? / I would like you to V., etc.).

Just like with indirect speech acts, multiple strategies can lead to the same target class of DF.

- enthusiastic agreement => refusal (EN: *like hell I will*, RU: aga ščas, kak že, ešče čego SR: *evo trčim, trčim pa se saplićem*)
- declaring the lack of need for an action => refusal (EN: *I'm good*, *No need*, RU: *ni k čemu*, SR: *nema potrebe, ne treba*)

The DF resulting from the two strategies differ in use: the first group are reactions to requests, and the second — reactions to offers.

The source **meaning** leaves a trace!

Pragmaticalization and variation

Does this variation mean that pragmaticalization differs from grammaticalization?

Despite this seemingly more significant flexibility, DF demonstrate the properties ascribed to grammaticalization:

- increase in frequency, context generalization
- syntactic reduction (parallel to phonetic erosion of grammaticalized lexical items), cf. RU *čto ty* [VP] (lit. 'what are you …'), SR *ma nemoj* [VP] (lit. 'don't …'), SL seveda [VP] (lit. 'it is known that …')
- semantic bleaching, cf. nontransparent RU *nado že* 'needed ptcl', SR *kamo sreće* 'where happiness-gen.sg', SL *v redu* 'in order'
- morphological decategorization of the parts, cf. RU *čto* / SR šta / SL *kaj* + V-2p, but not V-1p or V-3p)

Grammaticalization and variation

Additionally, variation is not characteristic of pragmaticalization only, it is found in grammaticalization as well:

- for a certain grammatical category, there is usually one purely grammatical marker
- however, besides it, there is a range of flexible constructions expressing the same grammatical meaning

cf. nominal plural in Russian:

stol-y 'table-nom.pl', dom-a 'house-nom.pl'…

VS kuča 'bunch' / more 'sea' / buket 'bouquet' / gora 'mountain' N-gen.pl

see Russian Constructicon (Bast et al. 2021, available at constructicon.github.io)

DF are numerous (> 100 in every language of our sample).

It is natural to have many ways to say 'yes' or 'no'.

Grammaticalization and pragmaticalization are similar phenomena, as they both go through the constructionalization process (see Bybee 2006, Noël 2007)

Expanding the language sample will help us to reveal new strategies of DF pragmaticalization and instances of lexical variation

...so, we invite you to contribute to our database!