Supplementary Information I: Data Preparation

January 15, 2024

Contents

1 Data basis

Our data basis includes Russian, Polish and Slovene as representatives of the main Slavic branches, for two historical stages, i.e. contemporary (ca. 1980–2021) and older (17th–19th c.). The data are taken from the corpora specified in Table 1.

Language	Historical stage	Corpus
Russian	contemporary	Russian National Corpus (RNC, NKRJa 2022)
	older	Russian National Corpus (RNC, NKRJa 2022)
Polish	contemporary	Polish National Corpus (PNC, NKJP 2012)
	older	Electronic Corpus of 17th–18th c. Polish Texts (KorBa, KorBa 2023) Diachronic Corpora of Polish (Diaspol)
Slovene	contemporary	Gigafida Corpus (Krek et al. 2019)
	older	IMP Korpus (Erjavec 2015)

Table 1: Corpora used for the extraction of data

For each language and period, we took random samples of 50 examples for the connectives COMP and DIR as single elements, and for the combination of COMP and DIR, applying the following criteria:

- COMP, DIR appear at left edge of clause
- combination of COMP and DIR, with maximal distance = 3 words
- DIR is not preceded by an argument
- DIR is not followed by a subjunctive particle (slv. *naj bi*, rus. *pust' by*, pol. *niech by*)
- DIR signals directive-optative illocutionary force (for other functions see Dobrušina 2019)

The combination of COMP and DIR is referred to as COMP-DIR as a cover term for adjacent and non-adjacent cases.

The data basis used for our analysis is given in Table 2; note that for pol_old, only 25 examples of COMP-DIR could be found on a non-random basis (i.e. all items found in three diachronic corpora were included).

2 Annotation

We annotated the examples for the variables and values given in Sections 2.1–2.21. Where necessary, we provide examples to illustrate our coding decisions.

		Tokens		
Sample	Corpus	COMP	DIR	COMP-DIR
pol_old	korba.edu.pl,	50	50	13
	diaspol.uw.edu.pl/XIX/#!			12
	diaspol.uw.edu.pl/polniem/#!			12
pol_cont	nkjp.pl	50	50	50
rus_old	ruscorpora.ru	50	50	50
rus_cont	ruscorpora.ru	50	50	50
slv_ old	nl.ijs.si/imp	50	50	50
slv_cont	viri.cjvt.si/gigafida	50	50	50

Table 2: Data basis: randomly extracted (with the exception of pol_old for COMP-DIR, see above) examples for COMP and DIR and the sequence of COMP and DIR for six samples specified by 'language+time'

2.1 LANG

Information on language.

pol Polish

rus Russian

slv Slovene

2.2 TIME

Information on temporal specification of sample.

old older (17th–19th c.)

cont contemporary (ca. 1980–2021)

2.3 LANGTIME

Combination of LANG and TIME, used to identify the six samples.

2.4 EX

Full example as taken from corpus.

2.5 CONN

Connectives relevant for our study.

COMP rus. *čto*, pol. *że*, slv. *da*

DIR rus. *pust'*, pol. *niech*, slv. *naj*

COMP-DIR rus. *čto pust'*, pol. *że niech*, slv. *da naj*

2.6 ADJ

Captures whether connectives are adjacent, relevant for COMP-DIR combinations only.

n COMP-DIR are not adjacent, see (1)

y COMP-DIR are adjacent, see (2)

dna 'does not apply', relevant for COMP / DIR

(1) non-adjacent COMP-DIR
Ему возразили, *что* лучше *пусть* не отмирают.
'They objected to him that they should better not die.' (rus_cont)

(2) adjacent COMP-DIR
Ali mi mogoče na glavi piše, *da naj* mi dajo "petdesetko"?
'Is it maybe written on my head that they should give me "fifty"?' (slv_cont)

2.7 INTER_CONN

Relevant for COMP-DIR combinations only. If ADJ='n' (Section 2.6), we give material intervening between connectives in terms of basic structural make-up, otherwise the specification is 'dna'. Note that we consider surface strings and do not imply any deeper syntactic analysis. We apply the same specifications for VFIN_ARG (Section 2.9).

dna does not apply (if ADJ='y')

adv adverb (including particles, such as Pol. *może* 'maybe', *aż* 'even', *owszem* 'well'), e.g. (3)

advp adverbial, e.g. (4)

dem demonstrative (including function as topicalizer)

S subject (including DAT- and GEN-subjects), e.g. (5)

- Sp: pronominal S (personal and demonstrative pronouns)
- Sc: clausal S
- Sa: appositive S
- Ssc: secondary S

DO Direct object, e.g. (6)

- DOp: pronominal DO (ACC, GEN), e.g. (7)

- DOc: clausal DO (including indirect questions)
- DOa: appositive DO

IO Indirect object (including ethical DAT, possessive DAT, free DAT, possessor raising), e.g. (8)

- IOp: pronominal IO

OblO Instrumental and prepositional objects (including goal/location for verbs of motion, e.g., 9)

S2 Nominal part of complex predicate

Examples

(3) INTER_CONN = ADV

Zawsze uważałem, że *po prostu* niech każdy myśli sobie, co chce …

'I've always thought that *just* let everyone think what they want …' (pol_cont)

(4) INTER_CONN = ADVP

to ja "na swój chłopski rozum" napiszę testament, że *po mojej śmierci* niech mieszkanie dzielą na połowę.

'so, "according to my peasant way of reasoning", I will write a testament that *after my death* let them divide the apartment in half.', or: '… they may divide…' (pol_cont)

(5) INTER_CONN = S

W związku z tym uważam, że *nazwa* niech zostanie…

'Therefore I think that *the name* may remain…', or: '…that may *the name* remain.'

(6) INTER_CONN = DO

Dlatego też odpowiem mu na konferencji prasowej, że *umoralnianie* to niech zostawi dla siebie.

lit. 'Therefore, I will answer him at the press conference that *moralizing* he should leave for himself.' (pol_cont)

(7) INTER_CONN = DOp

ampak takoj zjutraj je šel k mairju ter mu zapovedal, da *ga* naj prihodnjo nedeljo okliče

'but immediately in the morning he went to the major and ordered him, that he declares *it* next Sunday' (slv_old;)

(8) INTER_CONN = IO, with IO = ethical dative

gdy mu tłumaczę, że *swemu rozumowi* niech to przyzna, to za nic na to pozwolić niechce

'when I explain to him that he should admit it *to his reason*, he will not allow it for nothing' (pol_old)

(9) INTER_CONN = OBLO

Корабельников плюнул и сказал, что *с шаманами* пусть Мельников разговаривает rus.

'Korabel'nikov spat and said that *to the shamans*, Mel'nikov should talk' (rus_cont)

2.8 INTER_CONN_SYL

Material appearing between connectives, measured in syllables.

Counts Number of syllables

2.9 VFIN_ARG

Order of finite verb and arguments. By 'arguments' we mean nominal expressions filling valency slots of the verb; predicate instrumental and nominal parts of complex predicates are annotated as 'S2' (for details see Section 2.7).

V Finite verb (tense and number marking; no modal predicates lacking tense and person features)

dna No finite verb present, e.g. (12)

S Subject; first valency slot (including DAT-/GEN-subjects), e.g. (14)

DO Direct object; second valency slot

IO Indirect object; third valency slot

OblO Oblique object (OblOp if pronominal OblO), including prepositional objects

S2 Nominal part of complex predicate

Ssc Secondary subject, e.g. (17)

-p, -c, -a Further specification of arguments (pronominal, clausal, appositive)

Notes

- We do not consider clitic elements that do not specify arguments, e.g. rus. *бы*, pol. *się*. Such elements are considered elsewhere (INTER_CONN_VFIN, Section 2.11; INTER_CONN_VFIN_SYL, Section 2.12).
- We do not consider appositions, since they are not part of the argument structure. Appositions are considered elsewhere (INTER_CONN_VFIN, Section 2.11; inter_conn_Vfin_syl, Section 2.12).
- If the predicate and/or arguments are complex and their components are separated by intervening elements ('split' arguments), we consider the first component, e.g. (10), (11),
- Infinitives in control constructions are considered clausal direct objects (DOc), see (13).

Examples

(10) VFIN_ARG = S_OBLO_V, with split S

niech *dobrzy tobą cieszą się sąsiedzi.*

lit. 'May *good you enjoy neighbors*', i.e. 'may good neighbors enjoy you' (pol_old)

(11) VFIN_ARG = OBLO_V, with split OBLO

PIES z wielu cnot swoich notandus, niech też *do mego przyidzie Zwierzyńca.*

lit. 'The dog is notable for many virtues, may *it also to my come zoo.*' i.e. 'may it come to my zoo as well' (pol_old)

(12) VFIN_ARG = DNA

Считается, что в «идеале» каждый «двойник» способен оттянуть у основного кандидата от 5 до 7 процентов голосов.

'one assumes that ideally every double is able to take away from each candidate 5–7% of votes' (rus_cont)

(13) VFIN_ARG = S_V_DOC, with infinitive as DOC

Co Bóg złączy, niech człowiek nie waży się *rozłączać.*

'What God joins together, let man not dare to *separate* himself.' (pol_cont)

(14) VFIN_ARG = S__V_DO, with DAT-subject

Pierwej jednak niech *mi* wolno będzie przywołać projekcje pozostałych uczestniczek zgromadzenia i dokonać wzajemnej prezentacji.

'First, however, let me recall the projections of the other participants in the assembly and make a mutual presentation.' (pol_cont)

(15) VFIN_ARG = S2

иногда настолько прямо и ясно, что лучше пусть уж и остаются *обстоятельствами*, а не […]

'sometimes matters are as straightforward and clear that let there them remain circumstances and not […] ' (rus_cont)

(16) VFIN_ARG = IOP_V_DO

Ali mi mogoče na glavi piše, da naj *mi dajo "petdesetko"*?

'Is it maybe written on my head that they should give me "fifty"?' (slv_cont)

(17) VFIN_ARG = SP_SSC_V

пусть он сперва сам выпьет.

'let him drink first' (rus_cont)

2.10 VFIN_POSITION

Position of the finite verb in relation to the left context. VFIN corresponds to the verb considered in VFIN_ARG and INTER_CONN_VFIN_SYL.

I VFIN in initial position, i.e. INTER_CONN_VFIN_SYL = 0

C VFIN preceded by clitic, i.e. one-syllable word form (word form that does not independently carry accent) or personal pronouns, e.g. (18)

F VFIN in final position, i.e. no material at all following VFIN; verb-only clauses are annotated as VFIN_POSITION = I

M VFIN in medial position, i.e. there is material preceding and following VFIN, with preceding = more than one (clitic) element.

Notes

- I includes also verb-only clauses.
- C applies only if there is not more than one such element.
- F does not include DOC, discourse markers, and in general elements that are not part of the syntactic structure, i.e. V+ DOC / V+discourse marker = F.
- If VFIN appears after a clitic and there is no material following VFIN, this is annotated as C, according to the principle to consider the relation to the left context, see (19)

Examples

(18) VFIN_POSITION = C

Tu notri stoji zapisano: *Vsak* naj vzame svoj križ na rame.

'Inside it is written: *everybody* should take his own cross on his shoulder' (slv_old)

(19) VFIN_POSITION = C

smo zaprosili, naj *on* pove, kdaj bodo poslej praznovali občinski praznik.

'we asked him to tell [lit: *he* should tell], when the next community holiday will be' (slv_cont)

2.11 INTER_CONN_VFIN

Material intervening between (last) connective and finite verb, given in terms of linear sequence of syntactic units. Note that we consider surface strings and do not imply any deeper syntactic analysis. In addition to the items listed in Section 2.7, we consider the following elements:

none no unit occurring between (last) connective and finite verb

cl clause, e.g. (33)

par parenthetical expressions, e.g. (21)

qu quantifying expression, e.g. (23)

neg negation

refl (clitic) reflexive marker

agr person-number marker (with *l*-forms), e.g. Pol. *śmy*

Examples

(20) INTER_CONN_VFIN = CL

Dalej, niech teraz, *póki czas ma*, wszytko uprząta.

'Further, may he now, *while there is time*, tidy up everything.' (pol_old)

(21) INTER_CONN_VFIN = PAR

W podarunek niech ten — *rzecze* — w wieczny Pojdzie naszej przyjaźni braterskiej złocony Prezent szpaleru.

approx. 'Let this gift - *he says* - be included into the eternal golden line of our brotherly friendship.' (pol_old)

(22) INTER_CONN_VFIN = DEM

Jeśli świat pańskich porównań jest tak ograniczony, niech *tak* będzie.

'If the world of your comparisons is so limited, so be it.' (lit. '… may *so* it be') (pol_cont)

(23) INTER_CONN_VFIN = QU

Mówiłam: powiedzże mu, niech *tyle* nie pije.

literally 'I told you: do tell him, may *that much* he not drink.' (pol_cont)

(24) INTER_CONN_VFIN = INTER_CONN_VFIN = ADVP

imele izreči, kako da naj *po njunih željah* bode celjska glavna šola vravnana

'they had to say how according to their wishes Celje main school should be aligned' (slv_old)

2.12 INTER_CONN_VFIN_SYL

Material appearing between connective and finite verb, measured in syllables.

Counts Number of syllables

2.13 CTP_PRES

We specify whether there is a suitable complement taking predicate (CTP) present in the immediate context.

yli yes: left, immediately preceding (not necessarily adjacent, but in the preceding clause), e.g. (25)

ylf yes: left, farther away (i.e. not in immediately adjacent clause)

yr yes: to the right

n no, e.g. (26)

Notes

- The label CTP includes not only prototypical complement taking predicates (for a definition see Cristofaro 2023), but all predicates that may open up a semantic slot for a clause in the respective contexts.

Examples

(25) CTP_PRES = YLI

Tu notri stoji *zapisano*: Vsak naj vzame svoj križ na rame

'Dort drin steht geschrieben: jeder soll sein Kreuz auf die Schulter nehmen' (slv_old)

(26) CTP_PRES = N

jutri odpotujem, da se nikoli več ne vrnem v ta kraj

'I will leave tomorrow, such that I will never return to this place' (slv_old)

2.14 CTP_POS

Part of speech of CTP.

v verb, including participles

n noun, including verbal nouns in *Funktionsverbgefügen* and auxiliary-like structures as in (27).

pred predicative expression, i.e. copula + expression of state by means of neuter adjectives, adverbs etc., as in (30)

dem demonstrative, including *taki*, *tako* etc., see (28)

qu quantifying expression as in (29)

dna no CTP present, i.e. CTP_PRES = n (see Section 2.13)

Examples

(27) CTP_POS = N, with N as part of an auxiliary-like construction

Wybrałeś mnie choć *dawałem Ci do zrozumienia*, że "umarli niech odpoczywają w spokoju".

'You chose me even though *I let you know* that "let the dead rest in peace".' (pol_cont)

(28) CTP_POS = DEM

Kair to *taka* metropolia, że niech się Rzym schowa ile tam milionów mieszka.

'Cairo is *such* a metropolis that let Rome hide, so many millions live there.' (pol_cont)

(29) CTP_POS = QU

Wprowadził *tyle* zbędnego bełkotu do ontologii, że niech ręka boska broni.

'He introduced *so much* unnecessary gibberish into the ontology that may God's hand defend us.' (pol_cont)

(30) CTP_POS = PRED

Pewnie że niech mu tak Pon Bóg pobłogosławi!

'*Sure/Certainly* that may God bless him!' (pol_cont)

2.15 CTP_LEX

We provide the CTP in its citation form. The CTP heads the clause introduced by the connective. It may be of different parts of speech, which are annotated elsewhere (see section 2.14).

2.16 PRE_CONN_PRES

We annotate whether there is a clause-initial unit before the first connective with the clause containing the connective as the relevant research domain.

y yes

n no

2.17 PRE_CONN

If there is a unit preceding the first connective as defined in Section 2.16, we provide it in terms of basic structural make-up. We consider surface strings and do not imply any deeper syntactic analysis (see also Section 2.7).

The following values apply in addition to those used for INTER_CONN (Section 2.7) and INTER_CONN_VFIN (Section 2.11)

dna no unit preceding first connective, i.e. PRE_CONN_PRES = n

cjn connective, e.g. (31)

cl clause, e.g. (33)

wh wh pronouns

ptc particle, e.g. (34)

(31) PRE_CONN = CJN

ojciec mówi, że się wyprowadzi *i* że matka niech sama się z tym wszystkim pierdoli.

'my father says he's going to move out *and* that mother should f*** off with it all.' (pol_cont)

(32) PRE_CONN = PAR

Proszę, mości Marszałku Sejmowy, niech już ten projekt będzie przeczytany.

'*Please, sir, Marshal of the Sejm*, let this draft be already read.' (pol_cont)

(33) PRE_CONN = CL

No to tak uświadomiony, niech pan się natychmiast bierze do roboty.

'*Well, being aware of it* may you immediately get to work.' (pol_cont)

(34) PRE_CONN = PTC

ga je prekinila partijska Marička Ivanič, *češ* da naj skrajša in naj pove, kaj funkcionira

'he was interrupted by the chairwomen M.I. telling him to make it short and to tell how it works' (slv_cont)

2.18 COMMENT

Here we note any further observations that might be relevant, e.g.:

- conn-clause appears as quote
- possible CTP (if there is an element that may possibly be interpreted as CTP)
- complex constituent (in particular complex predicate), since in the annotation, we consider the leftmost element of complex constituents
- split constituent (if constituents are split)
- infinitive (if DOc = infinitive)

2.19 YEAR

Here we provide the year of the example as given in the corpus.

2.20 SOURCE

If available, we provide the source of the example as given in the corpus.

2.21 VPOS

The variable VPOS conflates VFIN_POSITION = C and VFIN_POSITION = I under the label CI, keeping the values M, F and DNA.

CI finite verb in initial position or preceded by a clitic / one syllable element

M finite verb in medial position

F finite verb in final positoin

DNA variable does not apply

For the feature specifications see Section 2.10.

References

Cristofaro, Sonia (2023). *Subordination*. Oxford: Oxford University Press.

Dobrušina, Nina (2019). "Status konstrukcij s časticami *pust'* i *puskaj* v russkom jazyke". In: *Russian Linguistics* 43 (1), pp. 1–17. DOI: 10.1007/s11185-018-09208-0.

Erjavec, Tomaž (2015). *Jezikovni viri starejše slovenščine. The IMP historical Slovene language resources*. URL: http://nl.ijs.si/imp.

KorBa (2023). *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)* URL: https://korba.edu.pl.

Krek, Simon et al. (2019). *Gigafida 2.0. Corpus of written standard Slovene*. Centre for Language Resources and Technologies, University of Ljubljana. URL: https://viri.cjvt.si/gigafida.

NKJP (2012). *Narodowy korpus języka polskiego*. URL: http://nkjp.pl.

NKRJa (2022). *Nacional'nyj korpus russkogo jazyka*. URL: www.ruscorpora.ru.