East Slavic parallel corpora: diachronic and diatopic variation in Belarusian, Ukrainian, and Russian

Dmitri Sitchinava
mitrius@gmail.com
Bilingual corpora

• Bilingual parallel corpora – contrastive linguistics, “small” typology (English vs. Russian, Czech vs. Slovene)
• Bilingual corpora can be symmetrical (Russian-English, English-Russian). The Norwegian team (HuNOR) calls only this symmetrical corpora “parallel”
• “Families” of bilingual corpora within some “mother corpora” (Czech, Russian National corpora, Norwegian, Lithuanian)
• Within the RNC: 15 languages parallel with Russian (Slavic, Germanic, Romance, Baltic, Armenian, Buryat, Estonian, Chinese); 70 million tokens
• Ukrainian/Russian and Belarusian/Russian – 9 million each
Ukrainian and Belarusian in parallel corpora

- Both Belarusian and Ukrainian are under-represented languages in the field of corpus linguistics.
- There exist no comprehensive national corpus for either
- The best existing monolingual corpora are, respectively, bnkorpus.info and mova.info
- The number of corpora-based research for them is also limited.
- Rather few Belarusian and/or Ukrainian texts are featured in the collections of massive parallel texts (Cysouw & Wälchli 2007) or multilingual parallel corpora. The Universal Dependencies corpora for B (translations from Russian, sometimes with mistakes) & U are rather small
(Post)-soviet translation between East Slavic: quality issues

- Machine translations (texts retrieved from the Internet), and even in the printed sources
- Looseness of translations (typical for most genres)
- Omissions (censure, just shortening etc.)
- Soviet era: Russianization; Post-soviet era: avoiding direct calques
Subnorms

• Both Belarusian and Ukrainian are languages with standard forms that were established relatively late.

• There still coexist multiple sub-norms in the written standards of either language, more “Russianized” and more “Westernized” ones, dating back to different political periods, 1930s vs. 1920s (a split clearly visible in Belarusian: *narkamaŭka* vs. *taraškevica* and less perceivable albeit existing also in Ukrainian).
Subnorms

• Due to the dialectal factors and the historical political divisions of the East Slavic territories there has existed a **diatopic variation** in the standard-oriented Ukrainian and Belarusian texts, reflecting both traditional dialects and local sub-norms, especially the Western Ukrainian sub-normative variant with less Russian (but more Polish and/or German) influence in both grammar and lexicon.
Russian *bylo*

- Modern Standard Russian has a construction derived from the Slavic Pluperfect, viz. the *bylo*-construction:
- an invariable particle *bylo* plus a form of past tense (finite or participial: *pošēl bylo* PF-go-PST.M.SG be-PST.N, *pošedšij bylo* PF-go-PARTCP.PST.M.SG.NOM be-PST.N).
- avertive
- cancelled attempt
- frame past
- With participles, it marks more often cancelled result
Unfinished action that is developed in a short span:

*I started reminding him of our appointment, but a dignified old lady in whom I recognized Madame Junker interrupted me saying it was her mistake.* [Vladimir Nabokov. *Look at the harlequins!* (1974)]

*Ja popytalsja bylo napomnit’ emu o našej dogovorennosti*<...> [S. Ilyin, 1999]

*L* PFV-try-PST.M.SG be-PST.N.SG PFV-remind-INF he-DAT about our-LOC.F.SG appointment-LOC.SG*
English counterparts

- Zero – 46% cases (Р было, but Q)
- *To be about to, just going to* – 12%
- Short span adverbial: *podumal bylo* PFV-think-PST.M.SG BYLO (for a moment), *pobežal bylo* PFV-run-PST.M.SG BYLO (took a few rapid steps), *načala bylo* begin.PFV-PST.F.SG BYLO (for a while) – 9%
- Mood: *would have +ed* – 7%
- *to try* – 7%
Eastern Slavic Pluperfect

• Until the 17th-18th centuries Russian used to have a Pluperfect construction with an inflected auxiliary that co-occurred only with finite past forms (pošël byl, byla, bylo, byli).
• The same more archaic construction, inherited from the Old East Slavic “supercompound” form with two auxiliaries, is still attested (and called Pluperfect, “anterior past”, or “remote past”):
  • (~standard) Ukrainian and Belarusian (cf. Xrakovskij 2015 or Sitchinava 2013)
  • some Russian dialects:
  • Northern Russian (cf. Pozharitskaya 1996, 2015):
  • Central dialects, eg the dialects of the Murom region (Ter-Avanesova 2016).
Semantic archaisms

- Usually more archaic than Modern Standard Russian *bylo* from the semantic point of view as well
- Allows for additional uses like frame past situation, cancelled result
  - *Dom sgorel byl, no ego otstroili*
  - House PF-burned.down-*PST.M.SG* be-*PST.M.SG* but it.M.ACC.SG PF-build-PST.PL
- ‘The house (lit. had) burned down, but it has been rebuilt since’
- Introduction marker in discourse (cf. residual use of the formula *žili-byli* ‘once upon a time, there lived’ in Standard Russian).
- These types of uses were also attested more or less in Old East Slavic (cf. Petrukhin, Sitchinava 2006) and are also known for Pluperfects cross-linguistically.
Pluperfect polysemy

(cf. Squartini 1999 on Germanic and Romance and further research)

• temporal precedence in the past
• past resultative
• closed temporal frames
• remoteness
• cancelled result (~25%, Dahl 1985)
• counter-factuality
• experiential uses
• evidentiality
• digression, backgrounding, marking initial fragments
Corpora-based study on Pluperfect distribution
Pluperfect in Europe
Pluperfect in Europe

• Consequence of tenses (SAE): most Germanic and Romance languages, **Sorbian**, Baltic Finnic or Latvian. Internal divergence is quite significant (e.g., in French Frame Past is marked rather by Imperfect; Scots or Hessisch use more Simple Pasts than the standard languages). (NB: Molise Slavic according to Barentsen)

• Less obligatory Pluperfects marking past resultatives or specially highlighting the consequence of events: under this label fall **Balkan Slavic** and Lithuanian (these properties correlate with those of rather “weak” Perfects in these languages; NB in Slavic Perfective aspect alone can mark anteriority)
Pluperfect in Europe

• Languages that use their (former) Pluperfects excessively rarely, mainly in residual contexts, viz. cancelled result or avertive (East Slavic like Rus. bylo) or irreality, usually together with Conditional byl by + / (West Slavic, Ukrainian, Belarusian and Slovene; in Conditional it is in fact a Past form)

• Turkish: marks all the digressions, states in the past, Frame Pasts, avertives (“I nearly died”, a rather rare function of Pluperfects)
Contexts

- The contexts that yield pluperfect in most European languages include the "iamitive" and reiterative contexts ('already', Ö. Dahl’s term). Cf. languages with "Weak" Pluperfects:
- "Many happy returns of the day," called out Pooh, forgetting that he had said it already.
- LT: - Širdingai linkiu tau viso labo!—šaukė Pūkuotukas, visai užmiršęs, kad šiandien jau buvo sakęs tą patį.
- ...be-PST.3SG say-PARTCP
- BE: – Zyču zdaroūja i radaści, – uskliknuū Pych, zabyūšysia, što jon užo pavinšavaū byū la raniej.
- ...PFV-congratulate-PST.M.SG be-PST.M.SG
- HR: - Moje iskrene želje za tvoj rođendan—dovikivao je, zaboravivši da je ovo već bio rekao.
- be-PST.M.SG say-PST.M.SG
Supercompound forms

- Based on a compound Perfect form (HAVE or BE + participle)
- The auxiliary is itself in compound Perfect > 2 auxiliaries
- Il est venu > il a été venu (standard French, dialects; Franco-Provençal)
- Ich habe gelesen > ich habe gelesen gehabt (colloquial)
- NB a uniformed « auxiliary of shift » in some languages with HAVE/BE auxiliary choice (Franco-Provençal, Yiddish)
Works on supercompounds

• Without typological generalizations until 1980s
• Holtus 1995 on Romance
• Litvinov, Radčenko 1998 about German with parallels
• Buchwald-Wargenau 2012 – German (diachrony)
• Gilbert Lazard 1996 – surcomposé on Iranian
• Lewin-Steinmann 2004 – Bulgarian and German
• Petrukhin et Sitchinava, 2006+ -- Slavic forms
• Europe mainly Romance & Germanic: Ammann 2005; Schaden 2009; L. De Saussure, Sthioul 2012
Areal distribution (roughly)
NB: Perfect vs. Past, areal
Russian language in Belarus: agreed
Pluperfect auxiliary

• *Na SSSR napali byli* (Minsk Radio)
• on USSR attack.PFV-PST.PL BE-PST.PL
• ‘The Soviet Union had been attacked’
• Perfect-in-the-Past
• *Stoilo mne bylo tol’ko podumat’, chto tebja moglo i ne byt’ v moej žizni...* (General Internet Corpus of Russian, Vitebsk)
• cost-PST.N.SG I.DAT be-PST.N.SG only think.INF that you-GEN may-PST.N.SG PART NOT BE-INF in my-LOC.F.SG life-LOC.SG
• As soon as I *thought* that you could have been absent in my life...
Russian language in Belarus: Agreed

- **EXPER:** “We once had an experience of...”
- A discussion of water leaks from neighboring property and resulting damage costs
- *Nas byli zatopili sosedi čerez ětaż*
- we.ACC BE-PST.PL PF-flood-PST.PL neighbor-PL.NOM through floor-SG.ACC
- “We have (lit. had had) once our flat flooded by neighbors who lived two floors upstairs”
Non-canonical Russian *bylo* in the parallel texts

• The non-canonical instances of *bylo* that are found in the translations of Belarusian fiction to Russian are of particular interest because they are not always directly transparent from the original (cf. the problem of “transparency” and “translationese” in Cysouw & Wälchli)

• Sometimes they emerge where in Belarusian there is no Pluperfect
Non-canonical Russian *bylo* in the parallel texts

- Vitaŭt Čaropka’s story with a trivial use of Bel Conditional:
  - *I mne xacelasja nešta sačynic’. Hetae nešta pačynalasja b slovami…*
  - *…begin-PST.N.SG COND…*
  - ‘And I wanted compose something; this something would begin like this…’

- Translation by Taccjana Zaryckaja
  - *Xotelos’ čto-to sočinit’. Ėto čto-to načinalos’ bylo slovami…*
  - *…begin-PST.N.SG be-PST.N.SG*

- A non-canonical *bylo* construction that has irreal semantics (attested for the Belarusian Pluperfect as well as typologically, cf. English counter-factual *If I had come*)
  - Russian *by*-Conditional, cognate to the Belarusian form, would be perfectly grammatical.
Non-canonical Russian *bylo*

- Cf. also Past Conditional in original texts (found also in colloquial Russian in Russia, Standard Polish and Ukrainian):
  - *Pereryla vse, gde ono tol’ko moglo bylo by byt’* (General Internet Corpus, Belarus)
  - PFV-dig-PST.F.SG everything where it-N.SG only can-PST.N.SG BE-PST.N.SG COND BE.INF
  - ‘(a certain woman) has searched all the places where it *could* possibly be’
Transparency

• Pierad vajennym pažaram jon pahareŭ byŭ jašče čysciej, navat i pahrebnika tady nie zastalosia. [Janka Bryl’, 1966]

• Do ètogo požara on pogorel bylo ešče počišče, daže I pogreba togda ne ostalos’ [translation by A. Ostrovsky]

• ‘Before that fire it had (already) burned down even more completely, without even cellar left’
Transparency/Non-standard *bylo* in the Russian language of Ukraine

- Comparable phenomena can be found also in translations from Ukrainian (including those made by bilingual Ukrainian-Russian writers).

- Išče *bulo up”jateryt’ podobalo* za takovoje zlodijanije [Hr. Kvytka, 1833]

- Ešče *bylo podobalo upjaterit’ za takovoe zlodejanie* [self-translated]

- ‘It *would have been necessary* to apply the punishment five times for such an evil deed’
Pluperfect: Diachronical dimension

• Decline of the frequencies of the (non-standard) Ukrainian Pluperfect in fiction (other than counterparts of *bylo*, and even these) towards the later Soviet period (100 > 60 ipm, only fiction)

• Revival with some Post-Soviet authors, but still rare
Pluperfect: Diatopical dimension

• Higher frequencies in the texts by the authors coming from the predominantly Ukrainian-speaking regions (NB the Center more than the West, although the West has “non-standard” uses) minus the North; in Belarus the non-standard uses are more characteristic for Western Belarus

• Pluperfect frequencies (ipm) on the General Regionally Annotated Corpus of Ukrainian (GRAC, courtesy of M. Shvedova / R. von Waldenfels)
Pluperfect: Diatopical dimension
Lexicon and standardization

- *Toska* ‘~yearning, nostalgia, misery, Angst’, a word with a high entropy of translation counterparts (e.g., 66 equivalents in Rus-En corpus, \( H = 1.6 \) for English, \( H = 0.6 \) for Ukrainian)
- Modern Ukrainian counterparts: *tuha, žurba, smutok*
- Higher entropy (\( H = 1.9 \)) for pre-1930 Ukrainian, more counterparts since defunct (cf. *žjel’* or *tusk* that are Western, “obsolete” *zanuda* or *toska* – cognate of Russian *toska*, avoided since 1930s as “too Russian”)
- Same tendency with Ukr. *čajka* ‘seagull’ (cf. Russian *čajka* ‘seagull’, Modern Standard Ukrainian *martyn*)
Syntax: Ukrainian animate-like accusative -a with body parts

• *prykusyty jazyk-a* lit. ‘bite tongue-GEN’ (‘to stop talking’) parallel to Russian *prikusit’ yazyk* with zero-marked ACC=NOM.INAN; some other phraseological units

• Fiction after 1990: ipm increases from 55,5 to 66,9 (exact Fisher test p< 0.00001)

• “Phraseologisation” (whereas some other semantic groups favoring -a such as “days and months” or “trees” shrink since the beginning of the 20th century)
Other topics of interest

• Active participles in -juč- (cf. Russian and Church Slavonic -jušč-, Polish -qc-) vs. “more Ukrainian/Belarusian” relative clauses; active in 1920s and decline since then; diatopically, present in particular along the borders

• Possessives like ixnij ‘their’ vs. indeclinable ix ‘oni.GEN’; ixnij absent in written Ukrainian until 1880s and standartized since, correlates (in Ukrainian and Belarusian) with concrete/abstract nouns (Bel. [ixny > ix] dom ‘their house’ but [ix > ixnaja] moc ‘their strength’), declines and is severely stigmatized in Russian as “illiterate” since 1930s

• Attenuative comparatives with po- (productive in Russia and rare in U & B)
Active participles in Ukrainian