sciendo
DOI 10.2478/jazcas-2023-0044

MULTIPLE INTERPRETATION AND FRAGMENTED TEXTS WITHIN A HISTORICAL CORPUS: THE CASE OF OLD EAST SLAVIC VERNACULAR WRITING

DMITRI SITCHINAVA
Department for Slavic Linguistics, Institute of Slavistics, Potsdam University, Potsdam, Germany

SITCHINAVA, Dmitri: Multiple Interpretation and Fragmented Texts within a Historical Corpus: The Case of Old East Slavic Vernacular Writing. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 266 – 274.

Abstract: The paper presents the issue of fragmented and/or ambiguously interpreted texts within the corpora of Old East Slavic vernacular writing. One of these corpora, the corpus of the Old East Slavic birchbark letters, is already available, the other, comprising the texts of Old East Slavic inscriptions, is under preparation. Due to the fragmentary state of many birchbark and epigraphy texts, their lemmatization and grammatical tagging may be uncertain and multiple interpretations may coexist. Some lemmas survive only in fragments which are nevertheless relevant for the study of lexicon. The grammatical status of many fragments may be firmly established despite lacking lexical information. However the relevant data on these fragments is not available in the word indices and corpora that take into consideration only best-preserved word forms. In the paper, the representation and annotation of such word forms within the Old East Slavic vernacular corpora is presented, and relative frequencies of such phenomena within the birchbark letter corpus are shown, with some case studies showing the relevance of the annotation of fragmented forms. The existing approaches, namely for the classical epigraphy within the EpiDoc standard and in the Hittite syntactic treebanks, are also briefly presented and compared to the solution found within the Old East Slavic vernacular corpora.

Keywords: lacunae, epigraphy, fragmented text, historical corpus, birchbark letters annotation, lemmatization, Old East Slavic

1 INTRODUCTION

An "ideal" linguistic corpus, including historical ones, typically requires unambiguous annotation of each token (including the cases of homonymy), although in practice this demand is often neglected. Nevertheless there exist specific texts where ambiguity is not removable, either as a part of their inherent design (which is the case of puns and some poetic texts) or as a consequence of the fact they lack an unquestionable interpretation – which is the case of many historical texts in extinct languages. This may arise due to the fact that the surviving readings are distorted by a copyist and call for a critical reconstruction that cannot be established for sure; *hapax legomena* and other factors of a text in a dead language may also contribute

into the situation. But yet another factor that is under scrutiny in the present paper are short texts that survive either without contexts or as fragments. This is known to be a property of many inscriptions (epigraphy monuments) as well as messages on a dedicated carrier found in archeological excavations, be it tablets of clay, papyri, or East Slavic birchbark letters. The fragmented states of (con)texts, sentences, and tokens, and the ambiguity of interpretations related thereto, are a challenge for building a linguistic corpus.

The paper is organized as follows. In Section 2 we present the state of the art of dealing with fragmented words and broken sentences, using the examples of the ancient epigraphy databases and Hittite corpus. In Section 3 we elaborate on the solutions of these issues as presented in the corpora of the Old East Slavic epigraphy and developed by the author. In the final section we propose discussion of the results.

2 STATE OF THE ART

A standard for building electronic database of epigraphy (mainly of Antiquity, with such languages as Ancient Greek, Latin, or Coptics) and papyri is being developed as a dialect of the TEI encoding standard. Known as EpiDoc, it is described in (Elliott et al. 2007–2020) and is used for building online repositories and databases for such sources as the Greek inscriptions of Northern Black Sea,[1] papyri,[2] the Vindolanda tablets from Roman Britain and others. The structure of the EpiDoc format provides for marking the lost or illegible passages and their probable size, annotating separate symbols interpreted uncertainly, as well as providing alternative lemmatization and analysis. The tradition of classical epigraphy studies includes extensive reconstructions of the lacunae, based on the precedent texts, formulas, or context. These reconstructions are typically included, with their hypothetical status marked, into the representation of the inscription. The interface of these databases typically involves switching between text-only and XML formats.

The lemmatization and even tokenization issues, however, are seen as secondary ones for the EpiDoc project, albeit certainly fully implementable within the TEI format. A separate subsection in the EpiDoc guidelines is "Tagging lexical words in the text and/or linking to lemmata for purposes of indexing or search".[3] This section states that "explicit markup of words (tokenization) and identification of their dictionary headwords (lemmatization) are both optional. Many projects simply leave these features unmarked, or rely on automated processes in search software to detect word-breaks and link to lemmatizing tools such as Morpheus". In this guideline, the task of lemmatization is related only to building lexical indices, whereas morphological or syntactical annotation is not mentioned at all. The developers are invited to link the

[1] https://iospe.kcl.ac.uk/corpus/index.html
[2] https://papyri.info/
[3] https://epidoc.stoa.org/gl/latest/idx-wordslemmata.html

lemmas of their corpus to an online lexicographical database, where the morphological information may be specified. The issue of alternative lemmas is discussed: a token that can be related to multiple lemmas due to a lacune may be marked as such. In the following example given by the authors, the fragment preceding a lacune may instantiate in Greek both the noun ἔκφευξις 'escape' and the verb ἐκφεύγω 'to flee'.

<w lemma="ἔκφευξις ἐκφεύγω" part="I">ΕΚΦΕΥΞ</w>
<gap reason="lost" extent="unknown" unit="character"/>

The EpiDoc guidelines do not mention cases when the set of alternative lemmas is not defined (say, if one annotates a fragmented rare proper name or a potential *hapax legomenon*).

In the paper (Molina – Molin 2016) the problem of reconstruction and broken clauses is addressed within the context of a syntactic corpus of Hittite.[4] The bulk of cuneiform tablets written in Hittite survived in a fragmented state. Unlike the databases of the EpiDoc family, in the Hittite corpus linguistics is in the focus, and, particularly, the syntactic level. The authors distinguish between cases when a "broken clause" can still be analyzed to some extent with regard to its syntactic structure and cases when a token is devoid of surviving syntactic context and this is informative only with regard to lexical and morphological statistics. Overall, five degrees of "brokenness" are specified in the annotation of Molina and Molin's corpus. A clause, the minimal unit of the Hittite corpus, is defined manually and then split automatically into tokens. The fragments of word forms that lack a reconstruction are dealt by the tokenizer either as separate tokens in their own right or as a part of a larger token (although the authors do not state it explicitly). At least the authors quote an example word form […]-*šarraš* glossed as nominative singular of the lemma *X-sarra.* In the paper alternative lemmas or morphological interpretations are not discussed.

3 OLD EAST SLAVIC VERNACULAR CORPORA: PROBLEMS AND SOLUTIONS

3.1 Properties of fragmented and ambiguous texts in the Old East Slavic writing

The Old East Slavic language is an extinct ancestor of Belarusian, Russian, and Ukrainian, corresponding to the period since the first attestations in the 11th century until the conventional date of 1400. For some descriptive purposes this latter date may be shifted.

The monuments of the Old East Slavic epigraphy (mainly on churches' walls or on archaeological objects) as well as birchbark letters, the medieval Slavic documents

[4] http://hittitecorpus.ru/

written on tree bark which was a cheap writing material in Eastern Europe, are the major classes of East Slavic texts that share many challenges with the corpora of cuneiform languages, classical papyri or the epigraphy of Antiquity. They are a valuable linguistic and historical source of everyday spoken and dialectal language, to a larger extent than literary and ecclesiastical texts that survive in the medieval East Slavic books. As such they are described as *non-bookish* or *vernacular* Slavic writing (Franklin 2002), a quality not always shared by the Ancient archaeological written sources (e.g. the Vindolanda tablets reflect pretty standard Latin around AD 100).

Much of the birchbark letters and inscriptions feature physical lacunae (often only a small fragment survives) or are not well legible, with different interpretations available for some symbols. As far as birchbark letters are concerned, 67% qualified either as "fragments" or "small fragments", whereas only 28% survived in full (according to the database gramoty.ru, see the next section in more detail). Compared to the cuneiform tablets or Greek inscriptions, the vernacular East Slavic lacunae are reconstructed with more difficulty. Unlike the Akkadian or Hittite cuneiform, different copies of the same text are rarely found in the Slavic inscriptions or birchbark letters; the exceptions are rare and related mainly to ecclesiastical texts. Unlike the classical Greek inscriptions, the Old East Slavic epigraphy is less formulaic. The epigraphical texts often have multiple interpretations on many levels: for example, for the oldest short Cyrillic inscription found in Old Rus' not less than five different interpretations with different transcriptions were proposed, without a convincing academic consensus (Medynceva 2000, pp. 21–31). As the Old East Slavic texts were normally written without spaces or other signals of word division (*scriptio continua*), competing versions of word division often coexist for different text chunks. Note that in many classical Greek and Roman inscriptions, as opposed to the books, the word segmentation was in place.

The most comprehensive description of the language of the Old East Slavic birchbark letters, the bulk of which is excavated in Veliky Novgorod in what is now Northern Russia is presented in the book (Zalizniak 2004). The book comprises a grammar sketch and a commented publication of the majority of the texts. It is completed with lexical indices (direct and *a tergo*) with extensive lists of word forms and their grammatical analysis. These indices make feel the degree of uncertainty in linguistic description of these texts: the sign "(?)", manifesting this incertainty, occurs in the direct index 642 times, the tag "possible" 313 times, the tag "unknown grammatical form or lexeme" 190 times, three dots marking fragments are attested 501 times. The overall size of the birchbark corpus, by the time the book was published, equalled 15,300 tokens and 3,200 lexemes (Zalizniak 2004, p. 15).

Within the main text of this description fragmented word forms relevant for the linguistic description are mentioned that did not find way into this index – exactly due to their fragmented status. For example, the word form that looks like *...avita*

(possibly ...[*pr*]*avita*) (birchbark letter 111, about 1240s–1260s) might belong to different verbal lexemes, ending in *...praviti* 'to direct, to settle' (maybe with a prefix) or just in *...aviti* (such as *staviti* 'to put', *izbaviti* 'to liberate' etc.). But whatever this lexeme would be, it is positively a verb in dual number and in second person, either imperative or present. This fact is important for the history of the Old Novgorod dialect, as this example is the latest attestation of dual number in verb, superceded by plural in East Slavic (Zalizniak 2004, p. 518). However, this word form cannot be found in the index (as there is no lexeme to ascribe it in an alphabetic list), and if an electronic corpus of birchbark letters was annotated by this index, the form would never be retrieved. Another example is the possessive construction *Vorontsja Vojkov prja…* '(a thing) belonging to Voronets Vojko' from the letter 332 (1160s–1170s). Syntactically it is a typical early Slavic possessive construction: the first name of this person is marked by genitive, the second name by the attributive adjective, followed by a noun starting with *prja…* This possessed object cannot be positively identified (Zalizniak 2004, p. 158; p. 432). Again, this noun is absent from the index, and in a corpus annotated by it, the whole construction "genitive + possessive + noun" would not be searchable.

3.2. Treatment of the fragments and ambiguity in the vernacular corpora

Special historical databases describing not only the texts but the materiality (photographs and description of the carriers, archeological data, etc.) of the birchbark letters[5] and inscriptions[6] are being developed by a Moscow epigraphical team; see (Sitchinava – Dyshkant 2021). In the epigraphy database the option that enables multiple interpretations, transcriptions and translations of the same inscription is entrenched inherently in the design. For example, on the page of the aforementioned 10th-century inscription five transcriptions and five interpretations are listed.[7]

These databases do not feature linguistic annotation. However, both are synchronized with respective subcorpora included into the Russian National Corpus (RNC, ruscorpora.ru). These are the corpora of birchbark letters and epigraphy (the latter is under development, due to be published in 2023); the programming of these corpora is made by Timofey Arkhangelsky and Anton Dyshkant using the indices of (Zalizniak 2004) with manual post-correction and additional annotation. For birchbark letters, a separate syntactic treebank is annotated on the base of the RNC subcorpus by Olga Lyashevskaya and is published in the Universal Dependencies repository as an independent project.[8] This treebank also meets some of the challenges described in (Molina – Molin 2016), although differs from them in that syntactic dependencies are marked there rather than constituents (Lyashevskaya, in print).

[5] http://gramoty.ru

[6] http://epigraphica.ru

[7] http://epigraphica.ru/epigraphy/inscription/show/177

[8] https://universaldependencies.org/treebanks/orv_birchbark/index.html

The specific challenges of the vernacular corpora are dealt with using the following approaches (all the examples provided from the birchbark corpus). Note that similar solutions may be applied to other epigraphical material as well, and some are in fact already implemented in other corpora, such as (1) in EpiDoc or (2) in Molina and Molin's Hittite project (see examples in Section 2 above).

1. Multiple variants for lemmatization and/or grammatical tagging of the same token (including partly surviving tokens) are provided:

<w addr="003:1"><ana lex="поклонъ" gr="S,m,sg,nom"></ana><ana lex="поклонъ" gr="S,m,sg,acc"></ana>поклонъ</w>
<w addr="003:1"><ana lex="Григша" gr="S,persn,m,sg,gen"></ana><ana lex="Гришка" gr="S,persn,m,sg,gen"></ana>грикши</w>

Within the same birchbark letter № 3 alternative analyses of the formulaic *poklonъ* 'bow, greetings' and the proper name *Grikši* are proposed. The first can be construed either as accusative (cf. Latin *salutem* 'greeting.ACC') or nominative, whereas the second one can refer to multiple lemmas: it might be either *Grigša* (with a devoiced consonant) or *Griška* (with misplaced letters) in a normativized form.

2. "Partly surviving" lemmas (see examples above). Lemmas with three dots (initial) are used in the annotation of the corpus. All the grammatical and dictionary information with regards to them is fully searchable.

<w addr="(332a):2a"><ana lex="Воронець" gr="S,persn,m,sg,gen"></ana>вороньцѧ</w>
<w addr="(332a):2a"><ana lex="Воиковъ" gr="A,possess,sg,m,acc"></ana>въ[и]къвъ</w>
<w addr="(332a):2a"><ana lex="прѧ…" gr="S,m,sg,acc"></ana>прѧ---</w>
</line>

This is the example addressed above with the unidentified possessed object with a name starting with *prja-*. It may be, however, accepted for sure that is a masculine noun in accusative singular within a certain construction.

<w addr="001:1"><ana lex="…ии" gr="S,persn,m,sg,dat,fragment"></ana>…[и]ю</w>

Here is an example of the (further unspecified) proper name ending in *-ii*.

Rare are the lemmas with lacunae both in the beginning and in the end (the grammatical status of this adjectival toponym *...oster...* is extracted from the context where multiple place names are listed in repeating constructions).

<w addr="001:1"><ana lex="…остер…" gr="S,topn,n,sg,gen,fragment"></ana>-остер…</w>

3. Fragments where segmentation into words is unclear. A dedicated tag "unknown_segmentation" is added there and the original *scriptio continua* may be preserved in difficult cases.

<w addr="414:3"><ana lex="въ" gr="incorrect,PR,gvrn:loc,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana><ana lex="вѣсъ" gr="incorrect,S,m,sg,loc,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana><ana lex="взвѣсити" gr="incorrect,V,partcp,act,non past,sg,m,nom,unknown_segmentation,lemma_unsure" correction="возвѣсѧ"></ana>вовѣсѧ</w>

In this example, the analysis may be either in two tokens *vo věsja* 'in the weight' or in a single token *vzvěsja* 'having weighted'.

4. Fragments with uncertain lemmas are marked with a dedicated tag "lemma_unsure" (see the example in 3. above).

5. Fully reconstructed tokens (with all the letters lost) are annotated on a par with surviving tokens, if included in a published transcription. This is an example from the same letters with multiple place names; here *selo* 'village' is reconstructed with certainty, as this document is heavily formulaic. Such tokens are marked by a dedicated tag "reconstr".

<w addr="001:1"><ana lex="село" gr="S,n,sg,gen,reconstr"></ana>(села)</w>

6. Fragmented tokens. The tokens that have considerable lacunae and are reconstructed with high probability are marked by a dedicated tag "fragment".

<w addr="036:1"><ana lex="гривна" gr="S,f,fragment"></ana>(гри)<lbr/>вна</w>

In this example the word *grivna* (a currency) is split by line break and only the half of the word survived (although, in this case the partial reconstruction is made with a good degree of certainty).

In Tab. 1 we give absolute and relative frequencies of the described phenomena in the annotation of the birchbark corpus. In the last four lines of the table the instances are given where multiple interpretations are ascribed to the same token, from two to five.

The share of texts where a phenomenon is attested is calculated twice: by all the documents of the birchbark corpus and only by the "good" documents that are not

smallest fragments or texts without interpretation (that is they have a significant amount of interpreted texts). In the last column we give the average number of instances of a given phenomenon per each text where they are attested at least once. This parameter is introduced because of differences between fully preserved documents and texts with lacunae (where we tend to attest more that one phenomenon of the kind).

	instances	share of tokens	texts	share of all the texts	share of "good" texts	instances per text where found
Ending-only lemmas	127	0.54%	95	7.83%	8.64%	1.34
Beginning-only lemmas	91	0.39%	71	5.85%	6.46%	1.28
Unknown segmentation	20	0.09%	19	1.57%	1.73%	1.05
Uncertain lemma	784	3.36%	465	38.30%	42.31%	1.69
Reconstruction (fully lost tokens)	264	1.13%	158	13.01%	14.38%	1.67
Fragments	1,134	4.86%	510	42.01%	46.41%	2.22
2 variants	1,756	7.53%	746	61.45%	67.88%	2.35
3 variants	113	0.48%	102	8.40%	9.28%	1.11
4 variants	38	0.16%	34	2.80%	3.09%	1.12
5 variants	5	0.02%	5	0.41%	0.45%	1.00

Tab. 1. Fragmented and ambiguous tokens within the Birchbark letter corpus

4 DISCUSSION AND CONCLUSIONS

The main contribution of the proposed approach is the filling of the gaps in annotation, providing tagging for the tokens that allow for a lexical or grammatical interpretation but are devoid of it under traditional annotation approach. This solution allows to search for syntactic constructions, for morphemes and derivational types that are present in the text.

The indices in (Zalizniak 2004), including the one *a tergo,* lack "ending-only" lemmas, that are quite well marked grammatically and often keep derivational suffixes as well. 127 lemmas of this kind is now annotated and feature in the treebank (Lyashevskaya, in print) as syntactic nodes.

The phenomena related to the fragmented state and ambiguity of texts are attested on a large scale. Ambiguous tokens are represented at least once in two thirds of birchbark letters, "broken" tokens and uncertain lemmas in 40% of cases. If ambiguity or fragmented word form is attested once in a given text, the average overall number is between two to three attestations per text. Fully reconstructed tokens are found more rarely because of relatively less successful reconstruction of birchbark letters as compared to the inscriptions of Antiquity (see above).

Specific markup such as "uncertain lemmas" and "unknown segmentation" which is not represented in the most widespread epigraphical standard is useful for lexical and syntactical analysis.

The implementation of these approaches in birchbark and epigraphy corpora can be further elaborated in problematic passages within "conventional" livresque literature in extinct languages.

References

Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, Ch., Vanderbilt, S. et al. (2007–2020). EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Accessible at: https://epidoc.stoa.org/gl/latest/.

Franklin, S. (2002). Writing, Society and Culture in Early Rus, c. 950–1300. Cambridge: Cambridge University press, XVI + 325 p.

Lyashevskaya, O. (in print). Syntactic annotation of the Old Russian Birchbark Letters Corpus (submitted).

Medynceva, A. A. (2000). Gramotnost' v Drevnej Rusi [Literacy in Old Rus']. Moskva: Nauka, 293 p.

Molina, M., and Molin, A. (2016). Syntactic Annotation for a Hittite Corpus: Problems and Principles. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016 (no pagination).

Sitchinava, D., and Dyshkant, A. (2021). Integration of the Old East Slavic Epigraphical Databases, Corpora and Indices. In Scripta & E-Scripta 21, pages 95–108.

Zalizniak, A. (2004). Drevnenovgorodskij dialekt [Old Novgorod dialect]. Moskva: Jazyki slavjanskoj kul'tury, 879 p.