Automatic alignment of the Psalterium Sinaiticum and the Septuagint Psalms

Abstract
This paper describes the work on automatically aligning the Psalterium Sinaiticum with the Septuagint psalms in the Tromsø Old Russian and OCS Treebank (TOROT). It briefly accounts for the transcription, text processing and manual annotation of the Psalterium Sinaiticum itself. It then explains the choice of Greek text, describes the automatic lemmatisation and morphological tagging of the Greek text and calculates and analyses the success rate in a small sample. Next the algorithm for automatic token-level alignment of texts is briefly described, and the success rate calculated and analysed. The results seem quite good from a quantitative perspective (over 90% accuracy in most cases), and it may seem tempting to try to use the data directly. However, a pilot study of aspect in the Greek and OCS text shows that the automatically processed Greek parallel leads to considerable data loss, and that much manual sifting of apparent mismatch examples is necessary to arrive at a preliminary analysis. In a low-resourced historical language such as Old Church Slavonic we cannot afford working with this amount of noise and data loss. We can use automatic tagging and alignment to ease our workload, but we have to manually post-correct the output.

1. Introduction

The overwhelming majority of texts in the Old Church Slavonic canon (however we define it) are translations from Greek. In many ways this is a limitation, since native Slavonic features cannot always be reliably distinguished from Greek influence. However, the existence of a Greek parallel may also be an asset for the study of certain grammatical phenomena. A case in point is the OCS aspect system, where the relatively well-understood Greek aspect system can often stand in for the lack of native speaker intuitions. When an Old Church Slavonic form, for example the past active participle, consistently translates a Greek aspectual form, namely the aorist active participle, and the two forms are formally dissimilar and also etymologically unrelated, we must assume that the identification is based on semantics, and any serious study of OCS aspect must take the Greek source text into account (see e.g. Dostál 1954, Amse-de Jong 1974, MacRobert 2013, Eckhoff & Haug 2015, Kamphuis 2020).

Any serious electronic corpus of Old Church Slavonic should therefore be a parallel corpus, insofar as a reasonably reliable Greek parallel text is available. In the PROIEL (Haug & Jøhndal 2008) and TOROT treebanks (Eckhoff & Berdicevskis 2015) we have made a start by providing the Codex Marianus (PROIEL) and the Codex Zographensis (TOROT) aligned at token level with the Greek Gospels (Tischendorf 1869–1872).[1] The obvious next step is to provide a fully annotated version of the Psalterium Sinaiticum aligned with the Septuagint

[1] Note that the Codex Zographensis is fully lemmatised and morphologically annotated, but that the syntactic annotation is only partial. Codex Marianus has complete syntactic annotation as well, as does the PROIEL version of the Greek New Testament.

Psalms, since the Greek source text is much less problematic than e.g. the source texts of the Codex Suprasliensis.

This paper describes the work on the TOROT parallel version of the Psalterium Sinaiticum. Section 2 describes the text processing and annotation of the Psalterium Sinaiticum in TOROT. Section 3 describes the choice and automatic processing of the Greek text, and evaluates the outcome. Section 4 describes and evaluates the automatic alignment of the Greek and OCS text. Section 5 is a pilot study which uses the obtained parallel corpus to replicate parts of Eckhoff & Haug's 2015 study of OCS and Greek aspect. Section 6 is the conclusion.

2. Psalterium Sinaiticum in TOROT

The version of the Psalterium Sinaiticum in TOROT is based on Severjanov's 1922 edition, and consequently contains Psalms 2–137. The text was manually typed,[2] slightly simplifying the Severjanov edition: all supralinear characters were taken down, and diacritics that were considered linguistically non-essential (primarily breathing marks) were omitted. All other features of the edition, including all of Severjanov's transcription choices, were retained.

The text was imported into TOROT's online annotation tool, which is an instance of the PROIEL Annotator web application.[3] It was automatically lemmatised and morphologically annotated, as described in Eckhoff & Berdicevskis 2015. An experienced annotator then manually corrected the lemmatisation and morphological analysis, and added manual syntactic annotation according to the PROIEL dependency scheme.[4] At the time of the automatic alignment and data extraction described in this paper, the annotation was not yet complete. Psalms 113–137 did not yet have syntactic annotation, but for the purposes of this paper, the lemmatisation and morphological of every verb had been checked and corrected manually. Only a small part of the text had been reviewed, i.e. checked by a second experienced annotator, which is a prerequisite for the annotated text to be released to the public on xml format.[5]

3. Choice and processing of Greek text

A corpus builder is often required to choose a particular text edition because it is the only one available or because it is a more convenient choice for some reason other than its quality alone. This was the case in our choice of edition for the Septuagint Psalms. The standard Septuagint text is the Göttingen critical edition, the most widespread text is the 1935 Rahlfs semi-critical edition (and the updated Rahlfs-Hanhart version), but neither of those are freely available. The latter also exists in a fully lemmatised and morphologically tagged version, but unfortunately it is so strictly licensed that it could not be used for our purposes. Instead, we chose to use Swete's diplomatic edition based on the Codex

[2] I am very grateful to Dr. Catherine Sykes for her hard work on the digitisation of this and many other texts in TOROT, as well as her excellent and perceptive work on the linguistic annotation.
[3] https://github.com/mlj/proiel-webapp
[4] For details about how the scheme was applied to Slavonic, see Eckhoff & Berdicevskis 2015, section 5.
[5] TOROT reviewed texts are released at https://torottreebank.github.io/

Vaticanus. This edition was the standard edition of the Septuagint for years after its publication, but has since been superseded by Rahlfs-Hanhart and the Göttingen critical edition. However, it is in the public domain and exists in a good and freely available digitisation,[6] and is therefore the best choice for our purposes. We were therefore obliged to provide lemmatisation and morphological tagging ourselves.

The text was automatically tagged and lemmatised using the standard import procedure used for all TOROT texts (Eckhoff and Berdicevskis 2015:11–12): the TnT Tagger (Trigrams 'n Tags, as described in Brants 2000) was trained on all the Greek materials in the PROIEL corpus (Haug and Jøhndal 2008, proiel.github.io), and the tagger was then applied to the Septuagint Psalms. Since the Greek text is already normalised, it was not necessary to apply any normalisation procedures, as we usually do for Slavonic texts. The tagger output was used in combination with direct lookups in the TOROT database (which contains the PROIEL tagged version of the Greek New Testament). For each word token in the texts, an algorithm checked whether that form already existed in the database. If it did, we assigned the morphological tag, part of speech and lemma of the existing form (or the most frequent combination if there were several hits). If not, the part of speech and morphological tag guessed by the tagger was assigned. A lemma was assigned by a simple lemma guesser. If the word form matched a lemma with the correct part of speech, that lemma was assigned. If not, characters were dropped from the end of the word form one by one, and the resulting string was checked again against the opening strings of lemmas of the correct part of speech. If a match was still not found, a dummy lemma ("FIXME") was assigned.

To improve the lemmatisation, we took advantage of the existence of publicly available lemmatisation for the Rahlfs 1935 text[7] (but unfortunately not for Swete). Since the Rahlfs text differs from the Swete text in many (usually minor) ways, we took a cautious approach. The algorithm worked verse by verse. Our lemmatisation was only checked against the Rahlfs lemmatisation if the verse had the same number of words in the two texts. 417 verses had mismatches, and the lemmatisation in those verses were thus left unchanged. For instance, Ps 1.5 has a word count of 13 in Swete but 12 in Rahlfs because Swete has an article that is missing in Rahlfs.

Next, the lemmatisation of the verse was checked word by word. If the lemma was the same in both Rahlfs and Swete, it was naturally retained. If the lemma was different, the following procedure was adopted:

If the automatically assigned lemma belonged to a list of known deviances between the Rahlfs lemmatisation policy and the PROIEL/TOROT lemmatisation policy, e.g. γίγνομαι versus γίνομαι, the assigned lemma was retained.

[6] https://github.com/eliranwong/LXX-Swete-1930/blob/master/01-Swete_word_with_punctuations.csv
[7] https://github.com/openscriptures/GreekResources/tree/master/LxxLemmas

If not, we checked if the Rahlfs lemma already existed in the TOROT database. If it did and this lemma had the part of speech tag automatically assigned to the word form, this lemma was assigned. If the lemma form did exist in the database, but with a different part of speech tag than assigned by the tagger, we assigned this lemma with the part of speech tag found in the database. Since the range of possible morphological tags differs from part of speech to part of speech, we had to discard the morphological tag provided by the tagger. Instead we assigned the tag "---------n" (non-inflecting) to all such word forms (767 such changes were made). If the word form in question was e.g. a preposition or a conjunction, this was in fact the correct tag, but it was also assigned to e.g. verbs (183 changes) and common nouns (169), which are rarely non-inflecting. The lemmatisation and part-of-speech tagging were thus given priority over the morphological tagging.

If the Rahlfs lemma was not found at all in the TOROT database, the automatically assigned lemma was retained (even if it was "FIXME"). After this procedure, 1876 word forms were still lemmatised as "FIXME".

To check the quality of the tagging and lemmatisation, Ps 138 (323 word tokens)[8] was manually corrected and compared with the automatic output. The results can be seen in Table 1.

Lemma + part of speech	Lemma form	Part of speech tag	Morphological tag
91%	92.6%	96.3%	84.5%

Table 1. Tagging and lemmatisation accuracy after enhanced lemmatisation

We can see that the lemmatisation and in particular the part of speech tagging was very good. 16 out of 24 lemma form errors were cases where the uncorrected text had "FIXME". The morphological tagging was weaker, at 84.5% accuracy. However, as seen in Table 2, in the majority of cases the morphology tag was off by only a single feature. Since the PROIEL/TOROT morphology tags are ten-place positional tags, we model this as Hamming distances.

0	1	2	4	5	6	7
273	38	6	1	2	1	2
84.5%	11.8%	1.9%	0.3%	0.6%	0.3%	0.6%

Table 2. Hamming distance between automatic and manual morphology tag

If we accept morphological tags with a single error, then, the success rate rises to 96.3%, the same as the part of speech tagging accuracy. When we analyse the errors in this group, we find that this is not an unreasonable thing to do. 16 of the errors are gender errors. As it turns out, 15 of these are cases where the tagger has suggested a supertag ("masculine or neuter", "masculine or feminine") for a form that can only be manually disambiguated, such as the personal pronoun σύ.[9] Only one of them has a real gender error (τρίβος was taken to be masculine rather than feminine by the tagger). 12 of the errors are case errors, all of them

[8] Deliberately chosen because Ps 137 is the last psalm in the Severjanov edition.

neuter accusatives that have been tagged as neuter nominatives or vice versa. Since these forms are always syncretic, this is a difficult task for the tagger. Four of the errors are examples of κύριος being tagged as an adjective rather than a common noun. The morphological tag is actually correct in all these four examples, but adjectives require one more feature than common nouns, namely degree, so all four examples have the tag 'p' for “positive” in the degree field in addition to the features the common noun had in the manual tagging. The final six errors all concern verbs: two examples of middle for passive voice, two examples of passive for active voice, one example of indicative for imperative, and one example of present for future tense.

4. Automatic alignment

The next step was to align the automatically tagged and lemmatised Greek text with the OCS text at token level. This was done with the alignment algorithm which is integrated in the PROIEL Annotator web application.[10] The algorithm is based on an automatic bilingual dictionary generated by ranking candidate lemmas in the source language on the basis of their likelihood of co-occurring in the same Bible verse as the target lemma. Candidate translation pairs within aligned sentences are then scored using the dictionary as well as the linearisation numbers within the sentence and the available morphological information.[11] Alignments that imply a transposition of word order are penalised (Haug et al. 2009, section 6).

An evaluation of the alignment of Psalms 2–94 (21,882 Greek word tokens in verses that also existed in the Psalterium Sinaiticum text) shows a success rate of 93.2%. Note that this also includes cases where no OCS word token was aligned with the Greek word token, which is overwhelmingly the case for definite articles, as seen in Table 3, which is an example of an automatically aligned verse with no errors. The two Greek articles (1 and 8 in the Greek linear order) are not aligned with any OCS token, just as the two OCS reflexive markers (5 and 10 in the OCS linear order) are not aligned with any Greek token. There is simply no correspondence.[12] The success rate is counted with Greek as the point of departure: the alignments of ὁ (1) and ὁ (8) with nothing are counted as correct alignment, but the lack of alignment of сѧ (5) and сѧ (10) is not counted at all. Ps 2.4 thus contains 11 correct token alignments.

ὁ	1		

[9] This followed from a weakness in the training data: PROIEL changed its annotation policy on personal pronouns midway through their annotation of the Greek New Testament, from no gender disambiguation to disambiguating gender if it was possible from context. The tagger thus had to deal with conflicting training data.

[10] https://github.com/mlj/proiel-webapp

[11] The original algorithm also used syntactic information, but since our Greek text has no syntactic tagging, the algorithm was adapted to run without it.

[12] Though of course the definiteness of ὁ (1) is rendered in the long form of Живъıі (1), and the reflexive content of сѧ (5) has a counterpart in the middle form of ἐκγελάσεται (5). However, the content is not isolated in a separate syntactic word.

κατοικῶν	2	Живъіі	1
ἐν	3	на	2
οὐρανοῖς	4	бсехъ	3
ἐκγελάσεται	5	посмѣетъ	4
αὐτούς	6	емоу	6
καὶ	7	і	7
ὁ	8		
κύριος	9	їъ	8
ἐκμυκτηριεῖ	10	поржгаетъ	9
αὐτούς	11	імъ	11
		сѧ	5
		сѧ	10

Table 3. Automatic alignment of Ps 2.4: "He that dwells in the heavens shall laugh them to scorn, and the Lord shall mock them."[13]

The alignment algorithm is better at getting token-with-token alignments right (95.9% correct) than token-with-zero alignments (84.1% correct).

An error analysis of the alignment of Psalms 2–9 and 54 yielded 103 alignment errors. 12 errors can be called partial – these were cases where a one-to-one alignment (which the Proiel Annotator requires) was not possible. An example is seen in Ps 5.10 (Table 4).

ὅτι	1	Ѣко	1
οὐκ	2	нѣстъ	2
ἔστιν	3		
ἐν	4	въ	3
τῷ	5		
στόματι	6	оустѣхъ	4
αὐτῶν	7	ихъ	5
ἀλήθεια	8	истинъі	6

Table 5. Automatic alignment of the opening of Ps 5.10: "For there is no truth in their mouth".

We see that Greek οὐκ ἔστιν (2, 3) is translated with нѣстъ (2) which is analysed as a single verb with incorporated negation.[14] In this case only one of the Greek word tokens can have an OCS alignment. The PROIEL/TOROT policy is to choose the word with the most lexical content in such cases, which would arguably be ἔστιν (3), but we see that the algorithm chooses οὐκ (2). The alignment in Table 5 therefore comes out with two errors in our count, one for οὐκ (2), which should have been unaligned, and one for ἔστιν (3), which should have been aligned with нѣстъ (2).

[13] The translations are mostly from Brenton's Septuagint translations, sometimes with some adjustments. The numbers in the second and fourth column indicate the linear order in Greek and OCS respectively.

[14] It would clearly have been possible to split нѣстъ into two syntactic word tokens, but it would be difficult to do so without distorting the form.

There were also two errors that were due to different verse boundaries in Swete and in the Severjanov edition of the Psalterium Sinaiticum. Since the verse is our unit of comparison, the algorithm cannot align tokens that belong to different verses.

The remaining 89 errors were all real and avoidable misalignments. In 44 of these cases the algorithm did not attempt an alignment at all, even though there was an obvious candidate in the corresponding OCS verse. This most frequently affected verbs (14 errors), common nouns (12 errors) and prepositions (10 errors). It is worth considering to what extent these errors could be due to the automatic tagging and lemmatisation.

All 14 verbs were correctly tagged as verbs, and had only minor errors in the morphological tagging. However, twelve of them were lemmatised as "FIXME". This clearly reduces the success of the collocation dictionary which is a central component of the alignment algorithm. An example can be seen in Ps 2.12 (Table 6). Although the OCS text has good alignment candidates for δράξασθε (1) and ὀργισθῇ (2) in the expected linear order, the algorithm is not able to make the alignment, probably because both Greek verbs are lemmatised as "FIXME" and the collocation dictionary does not suggest that this is a good alignment candidate for the two OCS verbs. Note that ἀπολεῖσθε (8), which is correctly lemmatised as ἀπόλλυμι, is successfully aligned with погъібнете (9)

δράξασθε (FIXME)	1		
παιδείας	2	наказанье	2
μή	3	когъда	4[15]
ποτε	4	еда	3
ὀργισθῇ (FIXME)	5		
Κύριος	6	їъ	7
καὶ	7	и	8
ἀπολεῖσθε (ἀπόλλυμι)	8	погъібнете	9
ἐξ	9	отъ	10
ὁδοῦ	10	пѫті	11
δικαίας	11	праведьна	12
		Приімѣте	1
		прогнѣваетъ	5
		сѧ	6

Table 6. Automatic alignment of Ps 2.12: "Accept correction, lest at any time the Lord be angry, and ye should perish from the righteous way".

We see a similar tendency with the prepositions, which seem to be particularly prone to "FIXME" lemmatisation, as well as part-of-speech errors (five out of ten errors had a preposition lemmatised as "FIXME" and tagged with an erroneous

[15] Note that this is also a misalignment: μή (3) should be aligned with еда (3) and ποτε (4) with когъда (4), just as the linear order suggests. This type of transposition error of frequently cooccurring pairs of words is quite common.

part-of-speech tag, two further errors had prepositions which were correctly lemmatised, but had the wrong part-of-speech tag). The tendency is less pronounced for the noun errors, where only four out of twelve errors had been lemmatised as “FIXME”.

In 43 cases the Greek word token was misaligned with an OCS token. In all of these cases there existed a correct OCS match, the error consisted in aligning the Greek word token with an OCS word token that either should have been aligned with a different Greek token (35 errors) or which should have remained unaligned (8 cases). This was most commonly seen with verbs (17 errors).

If we look at the errors involving Greek verbs, we again find that many of them have to do with lemmatisation problems. In six of the cases the verb was lemmatised as “FIXME”, and in four further cases the alignment appears to have been affected by other items in the same verse being lemmatised as “FIXME”. Both are exemplified in Ps 7.14 (Table 7), where the alignments of καιομένοις (καίω) and ἐξειργάσατο (FIXME) have been switched.

καὶ	1	И	1
ἐν	2	въ	2
αὐτῷ	3	немъ	3
ἡτοίμασεν (ἑτοιμάζω)	4	оуготова	4
σκεύη	5	съсѫдъі	5
θανάτου	6	съмрътънъїѧ	6
τὰ	7		
βέλη	8	Стрѣлъі	7
αὐτοῦ	9	своѧ	8
τοῖς	10		
καιομένοις (καίω)	11	съдѣла	10
ἐξειργάσατο (FIXME)	12	горѫщімъ	9

Table 7. Automatic alignment of Ps 7.14: “And in it he has prepared the instruments of death, he hath made ready his arrows for them that burn.”

However, we also find verb alignment errors in verses with no lemmatisation errors, such as in Ps 5.10 (Table 8).[16] Here we see that the correct alignment for the participle ἀνεῳγμένος (14) is the OCS adjective отвръстъ (12). Instead, the algorithm aligns it with естъ (10), which is actually the predicate of the preceding clause, and which has no counterpart in the Greek text.

ἡ	9		
καρδία	10	Срдцей	7
αὐτῶν	11	іхъ	8
ματαία	12	соуетъно	9

[16] See Table 5 for the opening of Ps 5.10.

τάφος	13	Гробъ	11
ἀνεῳγμένος	14	естъ	10
ὁ	15		
λάρυγξ	16	грътанi	13
αὐτῶν	17	ихъ	14
ταῖς	18		
γλώσσαις	19	ѣзъикъи	15
αὐτῶν	20	своими	16
ἐδολιοῦσαν	21	льщаахѫ	17
		отвръстъ	12

Table 8. Automatic alignment of Ps 5.10, second half: "their heart is vain; their throat is an open sepulchre; with their tongues they have used deceit".

We see, then, that the automatic alignment is generally of high quality, but it seems clear that the quality would have been considerably better if the tagging and particularly the lemmatisation of the Greek text had been better. It seems equally clear that even with a perfect Greek text some alignment errors would have remained. If we feel that we cannot afford losing datapoints (when the algorithm fails to align) or having alignment errors in our datasets, manual correction is necessary.

5. Aspect in the Psalterium Sinaiticum

To assess the value of our automatically added Greek parallel data, I will use it to replicate parts of Eckhoff & Haug's 2015 study on aspect in Marianus and Zographensis on the Psalterium Sinaiticum data. A long-standing dispute in the literature on Old Church Slavonic aspect is the question of whether the aorist and imperfect are really aspectual forms, or something else (see e.g. Dostál 1954:15–16, van Schooneveld 1951:96–97 and Meillet 1934:226–227 for attempts to define the aorist and imperfect as something other than exponents of viewpoint aspect). An important finding in Eckhoff & Haug 2015 is that the choice of aorist and imperfect closely follows the choice of aspect in the Greek gospel, and that verbs normally specialise with one inflectional aspect or the other. They also found that the modern prefix- and suffix-based derivational aspect formation was already largely grammaticalised, and could render Greek aspect in categories where no inflectional exponent was available.

To replicate parts of the study, a modified version of Eckhoff and Haug's query script was used to extract all verbs in the Psalterium Sinaiticum and their Greek automatic alignments (if any) from TOROT. This query yielded 5767 verb tokens. The lemmas *bųti, ne bųti* were then excluded due to their unusual behaviour, as in the Eckhoff & Haug study, yielding 5132 verbs, 4739 of which had a Greek alignment, 4625 of which had been automatically tagged as verbs.[17] A further 461 of these verb-tagged Greek tokens were lemmatised as 'FIXME', and 145 of those that were not were tagged as indeclinable.[18] Since we will look primarily at part-of-speech and morphological tagging, we can include the 'FIXME' tokens

[17] Note that many of the non-verb alignments may be correct, but they will be excluded from most of the statistics in this section because most of them require verbal morphological features (tense, mood) in the Greek aligned token.

tagged as verbs with a full verbal morphology tag,[19] which in practice leaves us with a dataset of 4480 word tokens. Thus, even before we start taking tagging errors into account, we have lost 652 datapoints, or 12.7% of the data.

Let us first examine to what extent aorists and imperfects are used to translate their Greek counterparts. Eckhoff & Haug 2015 measured this by creating a dataset consisting of *only* OCS aorists and imperfects translating Greek aorists and imperfects.[20] Their results are found in Table 9, the results from the Psalterium Sinaiticum are found in Table 10.

	Marianus		Zographensis	
	aorist	imperfect	aorist	imperfect
Greek aorist	98.6% (2887)	1.4% (42)	98.2% (2604)	1.8% (47)
Greek imperfect	11.1% (79)	88.9% (631)	11% (73)	89% (592)

Table 9. Translations of Greek aorists and imperfects, Marianus: n = 3639, Zographensis: n = 3316 (Eckhoff & Haug 2015:198)

	aorist	imperfect
Greek aorist	98.3% (1344)	1.7% (23)
Greek imperfect	32.9% (23)	67.1% (47)

Table 10. Psalterium Sinaiticum translations of Greek aorists and imperfects, n=1437

	aorist	imperfect
no verb morphology tag	3% (45)	6.7% (6)
aorist	89.9% (1344)	25.6% (23)
future	1.8% (27)	2.2% (2)
imperfect	1.5% (23)	52.2% (47)
pluperfect	0.1% (2)	1.1% (1)
present	2.5% (37)	11.1% (10)
perfect	1.1% (17)	1.1% (1)

[18] Recall from section 3 that this analysis was chosen when the Rahlfs lemmatisation suggested that the automatic part of speech tag was incorrect. These Greek tokens and their OCS alignments will also be excluded from most of the statistics in this section, since they are not tagged with verbal morphological features.

[19] But recall that such tokens often trigger alignment errors.

[20] While all imperfects and aorists in OCS are indicatives, this is not the case for Greek – the counts therefore also include Greek aorist imperatives, optatives, infinitives and subjunctives. Since the imperfective counterparts of these forms are traditionally called *present* imperatives, optatives, infinitives and subjunctives, and tagged accordingly, this slightly skews the data (1412 of the Greek forms were tagged as indicatives). To retain comparability, I did not change this.

Table 11. All Greek originals of Psalterium Sinaiticum aorist and imperfect translations, n=1585

We see that in all three datasets the aorist solidly translates the Greek aorist, at around 98% (Table 9 and 10). If we look at the full range of Greek tenses behind the OCS aorists and imperfects (Table 11), we see that the Greek aorist is indeed the main source of the OCS aorist translations at 89.9%.

The impression is strengthened when we look into the 23 occurrences where an aorist apparently translates a Greek imperfect: it turns out that 12 of the Greek verbs are mistagged aorists, while seven examples had either a Greek form or an OCS form which was ambiguous and could be either an aorist or an imperfect. There were thus only four certain discrepancies, such as example (1).

(1) Бъ же нашь на нб҃се ї на землі вьсѣ елико **вьсхотѣ** сътвори
ἐν τοῖς οὐρανοῖς καὶ ἐπὶ τῆς γῆς πάντα ὅσα **ἠβούλετο** ἐποίησεν
"But our God has done in heaven and on earth, whatsoever he has pleased." (Psalm 113:11, TOROT sentence ids 285434, 291882)[21]

In all these four cases the OCS translation seemed to reinterpret the Greek text and add a nuance which would require an aorist, in (1) an ingressive meaning which is indicated both with the prefix and the aorist form – 'decided' rather than stative 'wanted'. The deviation from the Greek is thus semantically motivated.

However, the distribution of imperfects appears to be significantly different in the Psalterium Sinaiticum[22] – the Greek imperfect is considerably less likely to be translated by an OCS imperfect – in fact 32.9% of them translate Greek aorists in this limited dataset (Table 10). If we look at the full range of sources for the imperfect (Table 11), we see that only 52.2% of the imperfects are translations of imperfects. The Greek aorist is the second most common source of the imperfect, at 25.6% (23 examples).

These 23 examples cannot be dispensed with as easily as the apparent imperfects translated as aorists. There are a few misannotations (five of the Greek verbs are misannotated imperfects) and ambiguous forms (five of the OCS verbs are at least technically ambiguous between aorist and imperfect, typically first-person singulars with stems in -a-), but 13 of the examples are unambiguous on both sides, as seen in example (2) and (3).

(2) **Расхыштахѫ** і вьсі мімоходѧшті пѫтемь-.. бъıстъ поношенію сѫсѣдомъ своімъ-..
διήρπασαν αὐτὸν πάντες οἱ διοδεύοντες ὁδόν ἐγενήθη ὄνειδος τοῖς γείτοσιν αὐτοῦ

[21] Severjanov's numbering is retained throughout. TOROT sentence ids are given both for the OCS and the Greek text, so that they can be directly retrieved at nestor.uit.no

[22] $p < 0.0001$, Fisher's Exact Test, compared to the Marianus distribution.

“All that pass by the way have robbed him: he is become a reproach to his neighbours.” (Psalm 88.42, TOROT sentence ids 284296, 291373)

(3) Егда **съкроушаахѫ сѧ** костi моѩ **поношаахѫ** ми врaѕи моi: Егда глѣаахѫ мнѣ на вьсѣкъ денъ къде естъ бг҃ъ твоi:-

ἐντῷ**καταθλάσαι** τὰ ὀστᾶ μου **ὠνείδισάν** με οἱ θλίβοντές με, ἐντῷλέγειναὐτούςμοικαθ’ἑκάστηνἡμέρανΠοῦἐστινὁ θεόςσου;

“While my bones were breaking, they that afflicted me reproached me; while they said to me daily, Where is thy God?” (Psalm 41.11, TOROT sentence ids 222679, 290355)

In (2), the OCS text seems to opt for a very reasonable telic-iterative interpretation (he is robbed repeatedly), while (3) seems to have a progressive reading, which we also see in the English translation (Brenton Septuagint Translation).

It may therefore seem that the Greek and OCS aorists are a better semantic match than the Greek and OCS imperfects. Nonetheless, all the deviant OCS imperfects appear to be semantically motivated – they are actively and creatively used to emphasise typical imperfect meanings, especially iterativity/habituality, even when the Greek text has an aorist. This strongly suggests that the aorist and imperfect are exponents of viewpoint aspect in the Psalterium Sinaiticum verb system, just as they are in the Marianus and Zographensis systems.

Another obvious observation to make is that the imperfect in general is much less frequent in the Psalterium Sinaiticum than in Marianus and Zographensis. This is not an effect of alignment problems: as seen in Table 12, only 12 imperfects remained unaligned.

	aorist	imperfect
all verbs	1618	102
verbs with Greek alignments	1495	90

Table 12. Number of aorists in the Psalterium Sinaiticum by alignment. p= 0.7665 (Fisher’s Exact Test)

It is more likely to be an effect of the contents of the Psalterium Sinaiticum – “poetic meditations on the relationship between God and his creation, which shift unpredictably and sometimes abruptly between narrative and appeal, between second and third person reference to the Deity, from past to present or future” (MacRobert 2013:397).

Given the very low share of imperfects, the aorist-imperfect contrast alone is of limited value as a diagnostic for the aspectual behaviour of individual verbs, and it therefore becomes especially important to see if we can also use the participle system for this purpose. Eckhoff & Haug 2015 show that there is a very strong correlation between Greek aspect and choice of participle form in Marianus and Zographensis.

	Marianus		Zographensis	
	past	present	past	present
aorist	98.8% (1070)	1.2% (13)	98.7% (938)	1.3% (12)
future	0	100% (2)	0	100% (1)
present	1.8% (23)	98.2 (1225)	1.9% (22)	98.1 (1109)
perfect	78.1 (178)	21.9 (50)	77% (157)	23% (47)

Table 13. OCS participles translating Greek participles, Marianus: n = 2561, Zographensis: n = 2286 (Eckhoff & Haug 2015:200)

	past	present
aorist	78% (39)	22% (11)
present	4.8% (21)	95.2% (413)
perfect	72.4% (42)	27.6% (16)

Table 14. Psalterium Sinaiticum participles translating Greek participles, n=542

In Table 14 we see that the correlation between the Greek and OCS present participle seems nearly as strong as in Marianus and Zographensis. The relationship between Greek perfect participles and OCS past participles is also about the same. However, the correlation between Greek aorist participles and OCS past participles appears to be significantly weaker.[23] However, when we examine the 11 apparent cases where a present participle translates a Greek aorist participle, we find that none of them are real: nine of them have a Greek present participle mistagged as aorist, and in the two final examples there turned out to be (manual!) annotation errors in the Psalterium Sinaiticum tagging – two past participles had been misannotated as present participles. When we look at the opposite group of mismatches – apparent translations of Greek present participles with OCS past participles, we see a similar picture: 18 out of 21 examples are due to mistagging of Greek perfect,[24] aorist or future participles as present participles, one is due to a (manual) mistagging of an OCS present participle as a past participle, and one is due to a misalignment. Only three examples are real discrepancies. In (4) and (5) the discrepancies are down to reasonable interpretation, while the less explicable (6) may suggest that Psalterium Sinaiticum may have been translated from a text with an aorist participle in this position.[25]

(4) Възмѣасѣа сѧ і вьсколѣбашѣа сѧ ѣко **пиѣні**: И вьсѣ мѫдрость іхъ поглъштена бъістъ: --

ἐταράχθησαν,ἐσαλεύθησανὡς ὁ **μεθύων**, καὶ πᾶσα ἡ σοφία αὐτῶν κατεπόθη.

“They are troubled, they stagger as a drunkard, and all their wisdom is swallowed up.” (Psalm 126.27, TOROT sentence ids 285206, 291757)
Greek is correctly tagged

[23] p<0.0001, Fisher’s Exact Test, compared with Marianus.
[24] There were 13 perfect participles in this group. Since the TnT algorithm uses *suffix* analysis for unknown words (Brants 2000:225), the stem-initial reduplication that signals Greek perfects is difficult to pick up.
[25] As the Codex Alexandrinus does.

The difference in (4) is clearly down to the fact that the Greek verb μεθύω means ‘to be drunk’, unlike OCS пити.

(5) **Сѣвъшеї** сльзами въ радость пожьнѫтъ:)-
οἱ **σπείροντες** ἐν δάκρυσινἐν ἀγαλλιάσειθεριοῦσιν.
“They that sow in tears shall reap in joy.” (Psalm 125.5, TOROT sentence ids 286026, 292219)

In (5), the past participle in OCS emphasises that the period of sowing must be finished before any joyful reaping can take place, which is a perfectly reasonable reinterpretation of the Greek sentence, even though the Greek present active participle does not carry any such meaning.

(6) **Створьшюмоу** чюдеса велиѣ единомоу:) Ѣко вь вѣкъ ми: -
τῷ **ποιοῦντι** θαυμάσια μεγάλα μόνῳ, ὅτι εἰς τὸν αἰῶνα τὸ ἔλεος αὐτοῦ·
“To him who alone has wrought[26] great wonders: for his mercy endures for ever.” (Psalm 135:5, TOROT sentence ids 289409, 292311)

We thus see that the correlation between Greek aorist participles and OCS past participles, and between Greek and OCS present participles in reality is just as strong in the Psalterium Sinaiticum as it is in Marianus and Zographensis.

We may note that the share of past participles (and not least the corresponding Greek aorist participles) is much lower than in Marianus and Zographensis. The reason for this is again probably the subject matter – past participles are typically a feature of narrative, as they are often used as so-called “conjunct participles”, which are often better translated as a coordinated full predicate than a subordinate adverbial modifier. To put it in the terms of Bary & Haug 2011, they are often independent rhemes rather than frames: they do not anchor the main verb in time or space, but independently drive the narrative forward. This is clearly seen in (7).

(7) Тъі **вьскресъ** помилоуеші сиона: Ѣко врѣмꙗ помиловаті: Ѣко приде врѣмꙗ:-
σὺ **ἀναστὰς** οἰκτειρήσεις τὴν Σειών, ὅτι καιρὸς τοῦ οἰκτειρῆσαι αὐτήν, ὅτι ἥκει καιρός
“Thou shalt arise, and have mercy upon Sion: for it is time to have mercy upon her, for the set time is come.” (Psalm 101.14, TOROT sentence ids 284741, 291577)

When we look at the most common syntactic uses of past and present participles in Marianus and the Psalterium Sinaiticum (Table 15), we see precisely this: While 61.5% of the past participles in Marianus are conjunct participles, only 11.4% of the past participles in the Psalterium Sinaiticum are. The numbers for present participles are similar (40.6% vs. 11%). In the Psalterium Sinaiticum,

[26] Here the Brenton Septuagint translation clearly reflects the aorist participle that we find in the Codex Alexandrinus, not the present participle of Codex Vaticanus.

participles are much more likely to be nominalised and turn up in argument position (subject, object, oblique argument) or to be adjectival and appear in attributive or predicative position.

	Marianus		Psalterium Sinaiticum	
	past	present	past	present
adverbial (ADV)[27]	3.4%	7.1%	3.6%	2.1%
apposition (APOS)	1.9%	4.4%	2.9%	5.1%
attributive modifier (ATR)	6.5%	10%	19.3%	24.1%
direct object (OBJ)	1.6%	2.7%	5.7%	6.4%
oblique argument (OBL)	2.2%	4.5%	5.7%	12.2%
predicate (PRED)[28]	0.5%	0.3%	6.4%	2.8%
subject (SUB)	4.5%	12.1%	7.9%	31.4%
conjunct participle (XADV)	61.5%	40.6%	11.4%	11%
argument with external subject (XOBJ)[29]	16.8%	17.3%	37.1%	3.9%

Table 15. Most common syntactic roles of participles in Marianus (n=2871)[30] and Psalterium Sinaiticum (n=576).[31] TOROT relation tags in parentheses.

If we are to use past-tense forms and participles as diagnostics of aspect in the Psalterium Sinaiticum, as almost all researchers trying to classify OCS verbs do, the analysis must therefore rely primarily on the aorist on the perfective side, but primarily on present participles on the imperfective side.

6. Conclusions

We have seen that we were able to perform high-quality automatic lemmatisation, part-of-speech and morphological tagging of the Septuagint Psalms, and also a similarly successful token-level alignment of the Greek text and the Psalterium Sinaiticum text. We have also seen that this automatically created parallel corpus enabled us to show that the correlation between Greek aspect and choice of OCS past-tense form and participle form is as strong in the Psalterium Sinaiticum as it was shown to be in Marianus and Zographensis, even though the distribution of past-tense forms and participles differs quite a lot between the Gospel texts and the Psalterium Sinaiticum.

[27] These are mostly dative absolutes.

[28] Mostly cases where a participle behaves like a main-clause verb, e.g. coordinated with another main-clause verb.

[29] Mostly passive participles in predicative position.

[30] Using Eckhoff & Haug's (2015) dataset, which is available at https://doi.org/10.18710/3YNHO7

[31] The syntactic annotation of the Psalterium Sinaiticum was not yet complete ate the time of data extraction (September 2020), these statistics use the subset of participles that had a syntactic relation tag.

However, we should not forget that this could not have been done without substantial manual inspection of apparent aspectual mismatches. Had we relied blindly on the numbers, we would have been left with a false impression that the Psalterium Sinaiticum translation was substantially less likely to follow the Greek aspect. We should also keep in mind that our relatively good results were achieved using a parallel corpus where the OCS text had high-quality manual annotation, while only the Greek text had automatic lemmatisation, part-of-speech tagging and morphological tagging, and where the alignment was performed automatically. If the OCS text had also been automatically lemmatised and tagged, we would probably have had difficulties coming to any firm conclusions at all.

The main problem with the automatic alignment was data loss – the verbs that the algorithm was unable to align simply had to be dropped from most of the statistics. The same was the case with part-of-speech tag errors: if a verb was not recognised as such, it was excluded from most of the statistics. Misalignments, on the other hand, seemed to cause very few problems. Errors and omissions in the lemmatisation did not have obvious direct consequences, but we know that they could cause alignment errors and thus data loss.

The morphological tagging had the lowest success rates of all the automatic processes, and predictably caused the most problems – a fairly large number of Greek aorist forms were regularly misinterpreted as imperfect or present forms and vice versa, creating noise in the material that had to be manually disentangled. We should note that in the present pilot study, these errors were only discovered where they created a mismatch with the OCS translation – it is equally possible that a number of real mismatches were overlooked because of similar tagging errors.

All in all, then, my view is that we cannot afford to use uncorrected automatic data of this kind. They may be good enough to give us an idea of major trends in the material, but the data loss and level of noise due to mistagging is more than we can afford in such a low-resourced language as OCS. The role of such methods should be restricted to preprocessing – they are good enough to save us a lot of work, but not good enough to provide data directly. The TOROT version of the Psalterium Sinaiticum will therefore be published with manually corrected alignments, and with the Greek lemmatisation, part-of-speech tags and morphological annotation manually checked. Only then will it be a genuinely useful tool in detailed studies of OCS aspect and other grammatical features.

References

Amse-de Jong, Tine (1974): *The meaning of the finite verb forms in the Old Church Slavonic Codex Suprasliensis: A synchronic study*. The Hague: Mouton.

Bary, Corien and Dag Haug (2011): Temporal anaphora across and inside sentences: The function of participles. *Semantics & Pragmatics* 4(8):1–56.

Brants, Thorsten (2000): TnT: a statistical partofspeech tagger. In S. Nirenburg (ed.): *Proceedings of the Sixth Conference on Applied Natural Language*

Processing 3, ANLC '00. Stroudsburg: Association for Computational Linguistics, 224–231.
Dostál, Antonin (1954): *Studie o vidovém systému v staroslověnštině*. Prague: Státní pedagogické nakladatelství.
Eckhoff, Hanne Martine & Aleksandrs Berdicevskis (2015). Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15.
Eckhoff, Hanne Martine & Dag T.T. Haug (2015): Aspect and prefixation in Old Church Slavonic. *Diachronica* 32(2), 186–230
Haug, Dag T. T. & Marius L. Jøhndal (2008): Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Caroline Sporleder and Kiril Ribarov (eds.). *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), 27–34.
Haug, Dag Trygve Truslew, Marius Jøhndal, Hanne Martine Eckhoff, Eirik Welo, Mari Johanne Bordal Hertzenberg and Angelika Müth (2009): Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of IndoEuropean Languages. *Traitement Automatique des Langues* 50.
Kamphuis, Jaap (2020): *Verbal aspect in Old Church Slavonic. A corpus approach.* Leiden: Brill Rodopi.
MacRobert, C.M (2013): The competing use of perfect and aorist tenses in Old Church Slavonic. *Slavia* 82(4), 387–407.
Severjanov, S.N. (1922): *Sinajskaja psaltyr'. Glagoličeskij pamjatnik XI v.* Petrograd: Rossijskaja akademija nauk.
Swete, Henry Barclay (1907): *The Old Testament in Greek according to the Septuagint.* Volume 2: *I Chronicles–Tobit.* Cambridge: Cambridge University Press.
Tischendorf, Constantin von (1869–1872): *Novum Testamentum Graece*. 8th edn. Leipzig: Hinrichs.