Lemmatization and POS-tagging

I observed lines that are not tokens themselves but are indicators of connection in following tokens from KTB test data. I got rid of them as they do not hold any comparative value, but their presence could lower the accuracy. For that, I used a regular expression, the following image shows what it picked up.

```
# text = 18 – 19 ғасырларда еуропалық өнердің ықпалымен сулы және майлы бояумен орындалған алғашқы туындылар өмірге келді.
8-9     сулы    _       _       _       _       _       _       _       _
11-12   майлы   _       _       _       _       _       _       _       _
# sent_id = Иран.tagged.txt:115:3455
# text = 20 ғасырда жаңа типті архитектуралық құрылыстар салына бастады.
# sent_id = Иран.tagged.txt:116:3474
```

```
Matching sentences count = 1029
Non-matching sentences count = 18
Total sentences count = 1047
```

There were 18 sentences not matching in token length between KTB and stanza annotations. Some observations:

- ’Қарама-қайшы’ is a single lemma in KTB while stanza broke it down into ’қарама’ and ’-қайшы’.
- ’Жұмысы жоқ’ is 2 tokens in KTB (noun and adjective respectively) and a single adjective in stanza.
- ’Лажсыз’ is 2 tokens in KTB (’лаж’ as a noun and ’сыз’ as an adposition) while it is a single-token adjective in stanza.
- Stanza ignores some words like ’көр’, ’жөн’, etc. It seems inconsistent as the same words appear in other sentences.
- Stanza turned one instance of ’-’ into a letter ’a’ and it was considered a part of the following token. All the other instances were okay, for instance ’-Жанейро’.
- Stanza did not break ’аман-есен’ into 3 separate tokens as KBT did. The opposite happened with ’әрең-әрең’, which Stanza considered 2 tokens (’әрең’ and ’-әрең’).

```
Шымкент – Қазақстандағы қала, Оңтүстік Қазақстан облысының орталығы.
Тұрғыны шамамен 683,273 адам (2014 жыл).
Қазақстанның басқа қалаларымен салыстырғанда тұрғыны жөнінен 3-ші орында (Алматы мен Астанадан кейін).
Осымен қатар, Шымкент Қазақстанның негізгі өнеркәсіп, сауда және мәдени орталықтарының бірі болып табылады.
Бірқатар археолог-ғалымдардың пікірлері бойынша, бұл жерде үлкен қорым болған, егер бұл растalса, онда қаланың пайда болған мерзімі қайта қаралуы мүмкін.
Қандай болғанда да, қала ескі заманнан-ақ адамдардың өмір сүруіне қолайлы мекен болған.
Оған ежелгі қоныстарға жүргізілген археологиялық қазба жұмыстары кезінде табылған мәдениет мұралары дәлел бола алады.
Тау етегінде өзен бойындағы алқаптарда егіншілік пен жүзімдік, ал көгалды таулы жайылымдарда-мал шаруашылығы дамыған.
1914 жылы Қазақстанның Ресей империясына қосылуының 50 жылдығына орай қалаға орыс генералы Черняевтің есімі берілді, бірақ 1921 жылы қала Шымкент атауына қайта ие болды.
XX ғасырда қала тарихының жаңа кезеңі басталды.
```

Figure 1: Text for Stanza.

1029 out of 1047 sentences matched in the number of tokens (98.2808%). Lemmatization accuracy is 98.7417%. POS tagging accuracy is 98.8951%.

- KBT rids tokens of suffixes as in ’жолғы’ which is evaluated as a noun after. It is intact and an adjective in Stanza. An adposition ’кейінгі’, for instance, should have been a verb ’кейін’.
- Stanza fails to recognize proper roots of words, for example, the root ’тақырып’ is ’тақырыб’.
- In Stanza ’мақтан’ is X, which is an adverb in KTB.
- KTB’s lemma for ’шығарушылық’ is ’шығару’, which makes sense given the sentence, but Stanza failed to recognize it and left the word as a whole.
- The lemma ’ешқандай’ is a denotation in Stanza and a pronoun in KTB.
- In KTB, the lemma ’жанкүйер’ is left as is, but Stanza found ’жанкү’ to be the lemma. ’әріптес’ is similarly turned into ’әріпте’.

```
Sentence accuracy = 98.28080229226362
Lemmatization accuracy = 98.74168797953963
POS tagging accuracy = 98.89514066496163
```

```
[{'lemma': 'осы', 'pos': 'DET'}, {'lemma': 'сектор', 'pos': 'NOUN'}, {'lemma': 'субъект', 'pos': 'NOUN'}, {'lemma': 'банкроттық', 'pos': 'NOUN'}, {'lemma': 'айқын', 'pos': 'ADJ'}, {'lemma': 'тетік
', 'pos': 'NOUN'}, {'lemma': 'енгіз', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': 'осы', 'pos': 'DET'}, {'lemma': 'сектор', 'pos': 'NOUN'}, {'lemma': 'субъект', 'pos': 'NOUN'}, {'lemma': 'банкроттық', 'pos': 'NOUN'}, {'lemma': 'айқын', 'pos': 'ADJ'}, {'lemma': 'тетік
', 'pos': 'NOUN'}, {'lemma': 'енгіз', 'pos': 'VERB'}, {'lemma': 'жөн', 'pos': 'ADJ'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': 'бұл', 'pos': 'PRON'}, {'lemma': 'біз', 'pos': 'PRON'}, {'lemma': 'заң', 'pos': 'NOUN'}, {'lemma': 'қарама', 'pos': 'ADJ'}, {'lemma': '-қайшы', 'pos': 'ADJ'}, {'lemma': '.', 'pos': 'PUN
CT'}]
[{'lemma': 'бұл', 'pos': 'PRON'}, {'lemma': 'біз', 'pos': 'PRON'}, {'lemma': 'заң', 'pos': 'NOUN'}, {'lemma': 'қарама-қайшы', 'pos': 'ADJ'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': 'нақ', 'pos': 'ADV'}, {'lemma': 'бұл', 'pos': 'PRON'}, {'lemma': 'мен', 'pos': 'PRON'}, {'lemma': 'әлем', 'pos': 'NOUN'}, {'lemma': 'ешбір', 'pos': 'DET'}, {'lemma': 'жерінен', 'pos': '
VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': 'нақ', 'pos': 'ADV'}, {'lemma': 'бұл', 'pos': 'PRON'}, {'lemma': 'мен', 'pos': 'PRON'}, {'lemma': 'әлем', 'pos': 'NOUN'}, {'lemma': 'ешбір', 'pos': 'DET'}, {'lemma': 'жер', 'pos': 'NOUN
'}, {'lemma': 'кер', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': 'православие', 'pos': 'NOUN'}, {'lemma': 'аҚазақстан', 'pos': 'NOUN'}, {'lemma': 'ең', 'pos': 'ADV'}, {'lemma': 'көп', 'pos': 'ADJ'}, {'lemma': 'тара', 'pos': 'VERB'}, {'lemma': 'конфес
сия', 'pos': 'NOUN'}, {'lemma': 'бір', 'pos': 'NUM'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': 'православие', 'pos': 'NOUN'}, {'lemma': '–', 'pos': 'PUNCT'}, {'lemma': 'Қазақстан', 'pos': 'PROPN'}, {'lemma': 'ең', 'pos': 'ADV'}, {'lemma': 'көп', 'pos': 'ADJ'}, {'lemma': 'тара', '
pos': 'VERB'}, {'lemma': 'конфессия', 'pos': 'NOUN'}, {'lemma': 'бір', 'pos': 'NUM'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': 'адам', 'pos': 'NOUN'}, {'lemma': 'лажсыз', 'pos': 'ADJ'}, {'lemma': 'ел', 'pos': 'NOUN'}, {'lemma': 'тілек', 'pos': 'NOUN'}, {'lemma': 'бағын', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'P
UNCT'}]
[{'lemma': 'адам', 'pos': 'NOUN'}, {'lemma': 'лаж', 'pos': 'NOUN'}, {'lemma': 'сыз', 'pos': 'ADP'}, {'lemma': 'ел', 'pos': 'NOUN'}, {'lemma': 'тілек', 'pos': 'NOUN'}, {'lemma': 'бағын', 'pos': 'VE
RB'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': 'сен', 'pos': 'PRON'}, {'lemma': 'бұл', 'pos': 'DET'}, {'lemma': 'жол', 'pos': 'NOUN'}, {'lemma': 'аман-есен', 'pos': 'ADV'}, {'lemma': 'қайт', 'pos': 'VERB'}, {'lemma': ',', 'pos': 'PU
NCT'}, {'lemma': 'ел', 'pos': 'NOUN'}, {'lemma': 'аман', 'pos': 'ADJ'}, {'lemma': 'бар', 'pos': 'VERB'}, {'lemma': '...', 'pos': 'PUNCT'}, {'lemma': '—', 'pos': 'PUNCT'}, {'lemma': 'де', 'pos': 'V
ERB'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': 'сен', 'pos': 'PRON'}, {'lemma': 'бұл', 'pos': 'DET'}, {'lemma': 'жол', 'pos': 'NOUN'}, {'lemma': 'аман', 'pos': 'ADJ'}, {'lemma': '–', 'pos': 'PUNCT'}, {'lemma': 'есен', 'pos': 'ADJ'},
 {'lemma': 'қайт', 'pos': 'VERB'}, {'lemma': ',', 'pos': 'PUNCT'}, {'lemma': 'ел', 'pos': 'NOUN'}, {'lemma': 'аман', 'pos': 'ADJ'}, {'lemma': 'бар', 'pos': 'VERB'}, {'lemma': '...', 'pos': 'PUNCT'
}, {'lemma': '—', 'pos': 'PUNCT'}, {'lemma': 'де', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': '"', 'pos': 'PUNCT'}, {'lemma': 'ол', 'pos': 'PRON'}, {'lemma': 'біз', 'pos': 'PRON'}, {'lemma': 'біл', 'pos': 'VERB'}, {'lemma': '!', 'pos': 'PUNCT'}, {'lemma': '"', 'pos': 'PUNCT'}, {
'lemma': 'де', 'pos': 'VERB'}, {'lemma': 'ой', 'pos': 'NOUN'}, {'lemma': 'кел', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': '"', 'pos': 'PUNCT'}, {'lemma': 'ол', 'pos': 'PRON'}, {'lemma': 'біз', 'pos': 'PRON'}, {'lemma': 'біл', 'pos': 'VERB'}, {'lemma': 'ау', 'pos': 'PART'}, {'lemma': '!', 'pos': 'PUNCT'}, {
'lemma': '"', 'pos': 'PUNCT'}, {'lemma': 'де', 'pos': 'VERB'}, {'lemma': 'ой', 'pos': 'NOUN'}, {'lemma': 'кел', 'pos': 'VERB'}, {'lemma': '.', 'pos': 'PUNCT'}]


[{'lemma': '(', 'pos': 'PUNCT'}, {'lemma': 'әлдеқашан', 'pos': 'ADV'}, {'lemma': 'жібер', 'pos': 'VERB'}, {'lemma': 'қой', 'pos': 'VERB'}, {'lemma': 'арыз', 'pos': 'NOUN'}, {'lemma': 'е', 'pos': '
AUX'}, {'lemma': 'жұмысы жоқ', 'pos': 'ADJ'}, {'lemma': ')', 'pos': 'PUNCT'}, {'lemma': '.', 'pos': 'PUNCT'}]
[{'lemma': '(', 'pos': 'PUNCT'}, {'lemma': 'әлдеқашан', 'pos': 'ADV'}, {'lemma': 'жібер', 'pos': 'VERB'}, {'lemma': 'қой', 'pos': 'VERB'}, {'lemma': 'арыз', 'pos': 'NOUN'}, {'lemma': 'е', 'pos': '
AUX'}, {'lemma': 'жұмыс', 'pos': 'NOUN'}, {'lemma': 'жоқ', 'pos': 'ADJ'}, {'lemma': ')', 'pos': 'PUNCT'}, {'lemma': '.', 'pos': 'PUNCT'}]
```

Figure 2: Sentences with different number of words. Stanza followed by KTB.

Named Entity Recognition (NER)

Tokens that were found not to be named entities receive 0, which seems to cause a zero-division error when running classification_report on a token that is evaluated as 0 by either of the two sets. It made sense to treat such cases as a difference, so I set those to 0 in the function parameters. However, it did not seem to have changed the accuracy, so that must be the default.
I also changed S to B and I to E before running seqeval functions. Without these changes, comparisons would not run. IOBES is supported in seqeval, but automatically downgrading one dataset into sensible alternatives to match the other made more sense as opposed to manually fixing the preprocessed kazNERD.

The accuracy is 98.788%.

- 'Қазақстан халқына Жолдауында' was found to be project name, but are not entity names in kazNERD, except for the name of the country, which is a geopolitical entity. Stanza might have done a better evaluation.
- 'Арқалық ет комбинаты' is an organisation name in kazNERD, but a facility in Stanza, again, Stanza's choice is close if not closer.
- 'еліміздегі Төтенше және өкілетті елшісі' yields very similar results - position for each token, but Stanza sees 'еліміздегі' as part of it, which is wrong.
- 'Америка Құрама Штаттарының' is a position in Stanza and a geopolitical entity in kazNERD, the latter seems like a better evaluation.
- 'In осы айдың алдында' both see a date, but Stanza excludes the third token.
- 'кем дегенде тағы 1 жыл' consists entirely of date for kazNERD, while Stanza leaves off first 3 tokens. both see a date, but Stanza excludes the third token.

```
Бірақ Қасым-Жомарт Кемелұлы ашық адам ретінде әртүрлі пікірлерді естуге дайын тұрады .
Сонымен бізге тікелей қатысты өте қуаныштысы — ол мына Семей қаласын тарихи орталық ретінде тану .
Небәрі үш медбике мен бір жалпы тәжірибелі дәрігер екі мың жарымға жуық адамның саулығына жауапты .
Небәрі 3 медбике мен 1 жалпы тәжірибелі дәрігер 2 , 5 мыңға жуық адамның саулығына жауапты .
Келесі кезеңде біз ( i ) елу ( ii ) тұрғыны бар ауылдарды кең жолақты интернетпен ( ii ) қамтимыз .
Келесі кезеңде біз ( i ) 50 ( ii ) тұрғыны бар ауылдарды кеңжолақты интернетпен ( ii ) қамтимыз .
Отбасылық құндылықтарды қадірлеу , ата-баба салтын ұмытпау керектігін ұғындырды .
Бұл — Үкіметтің тапсырмасы .
```

Figure 3: Text for Stanza

```
22 O O
23 O O
24 O O
25 O O
26 B-POSITION B-POSITION
27 I-POSITION I-POSITION
28 B-GPE B-PROJECT
29 O I-PROJECT
30 O I-PROJECT
31 O O


0 O O
1 B-POSITION B-POSITION
2 O O
3 B-DATE B-DATE
4 I-DATE I-DATE
5 I-DATE I-DATE
6 O O
7 O O
```

Figure 4: Token level comparison - kazNERD first, Stanza second

```
, 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'B-POSITION', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'],
['O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['B-ORGANISAT
ION', 'I-ORGANISATION', 'I-ORGANISATION', 'I-ORGANISATION', 'I-ORGANISATION', 'O', 'O', 'B-PROJECT', 'I-PROJECT', 'I-PROJECT', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O
', 'B-GPE', 'I-GPE', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['B-DATE', 'O', 'O', 'B-NORP', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'B-PRODUCT
', 'I-PRODUCT', 'I-PRODUCT', 'I-PRODUCT', 'O', 'O', 'O', 'O', 'O', 'O', 'B-QUANTITY', 'I-QUANTITY', 'I-QUANTITY', 'I-QUANTITY', 'O', 'O'], ['O', 'B-PRODUCT', 'O', 'O', 'O', 'O', 'O', 'O', 'B-QUANT
ITY', 'I-QUANTITY', 'O', 'O'], ['O', 'B-TIME', 'I-TIME', 'I-TIME', 'I-TIME', 'B-CARDINAL', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'B-TIME', 'I-TIME', 'I-TIME', 'B-CARDINAL', 'O', 'O', 'O', 'O', 'O', 'O'], ['O
', 'B-POSITION', 'O', 'B-DATE', 'I-DATE', 'I-DATE', 'I-DATE', 'I-DATE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-POSITION', 'I-POSITION', 'B-PRO
JECT', 'I-PROJECT', 'I-PROJECT', 'O'], ['O', 'B-POSITION', 'O', 'B-DATE', 'I-DATE', 'I-DATE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-POSITION'
, 'I-POSITION', 'B-PROJECT', 'I-PROJECT', 'I-PROJECT', 'O'], ['O', 'B-PERSON', 'I-PERSON', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-GPE', 'O',
'O', 'O', 'O', 'O', 'O'], ['O', 'B-CARDINAL', 'O', 'O', 'B-CARDINAL', 'O', 'O', 'O', 'B-CARDINAL', 'I-CARDINAL', 'I-CARDINAL', 'I-CARDINAL', 'O', 'O', 'O', 'O'], ['O', 'B-CARDINAL', 'O', 'O', 'B-
CARDINAL', 'O', 'O', 'O', 'B-CARDINAL', 'I-CARDINAL', 'I-CARDINAL', 'I-CARDINAL', 'I-CARDINAL', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'B-CARDINAL', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-CARDINAL', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O
'], ['O', 'O', 'B-ORGANISATION', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O'], ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]
```

Figure 5: NER results

```
Accuracy = 0.9878758935173774
              precision    recall  f1-score   support

       ADAGE       0.62      0.26      0.37        19
         ART       0.90      0.95      0.93       229
    CARDINAL       0.97      0.98      0.97      2824
     CONTACT       0.85      0.85      0.85        20
        DATE       0.96      0.97      0.97      2611
     DISEASE       0.94      0.89      0.92       123
       EVENT       0.86      0.83      0.84       156
    FACILITY       0.74      0.70      0.72       198
         GPE       0.97      0.96      0.97      1742
    LANGUAGE       1.00      0.93      0.97        46
         LAW       0.63      0.62      0.62        55
    LOCATION       0.84      0.84      0.84       212
MISCELLANEOUS       0.91      0.77      0.83        26
       MONEY       0.98      0.99      0.99       441
   NON_HUMAN       0.00      0.00      0.00         1
        NORP       0.98      0.88      0.93       372
     ORDINAL       0.97      0.94      0.95       386
ORGANISATION       0.85      0.89      0.87       735
  PERCENTAGE       0.98      0.98      0.98       456
      PERSON       0.97      0.98      0.97      1332
    POSITION       0.96      0.97      0.97       603
     PRODUCT       0.79      0.68      0.74        73
     PROJECT       0.91      0.91      0.91       211
    QUANTITY       0.95      0.95      0.95       407
        TIME       0.93      0.92      0.93       230

   micro avg       0.95      0.95      0.95     13508
   macro avg       0.86      0.83      0.84     13508
weighted avg       0.95      0.95      0.95     13508
```

Figure 6: NER comparison results

Entity	**Precision**	**Recall**	**F1-Score**	**Support**
ADAGE	0.62	0.26	0.37	19
ART	0.90	0.95	0.93	229
CARDINAL	0.97	0.98	0.97	2824
CONTACT	0.85	0.85	0.85	20
DATE	0.96	0.97	0.97	2611
DISEASE	0.94	0.89	0.92	123
EVENT	0.86	0.83	0.84	156
FACILITY	0.74	0.70	0.72	198
GPE	0.97	0.96	0.97	1742
LANGUAGE	1.00	0.93	0.97	46
LAW	0.63	0.62	0.62	55
LOCATION	0.84	0.84	0.84	212
MISCELLANEOUS	0.91	0.77	0.83	26
MONEY	0.98	0.99	0.99	441
NON_HUMAN	0.00	0.00	0.00	1
NORP	0.98	0.88	0.93	372
ORDINAL	0.97	0.94	0.95	386
ORGANISATION	0.85	0.89	0.87	735
PERCENTAGE	0.98	0.98	0.98	456
PERSON	0.97	0.98	0.97	1332
POSITION	0.96	0.97	0.97	603
PRODUCT	0.79	0.68	0.74	73
PROJECT	0.91	0.91	0.91	211
QUANTITY	0.95	0.95	0.95	407
TIME	0.93	0.92	0.93	230
Micro Avg	0.95	0.95	0.95	13508
Macro Avg	0.86	0.83	0.84	13508
Weighted Avg	0.95	0.95	0.95	13508

Table 1: NER Evaluation Metrics

Stanza on fiction text

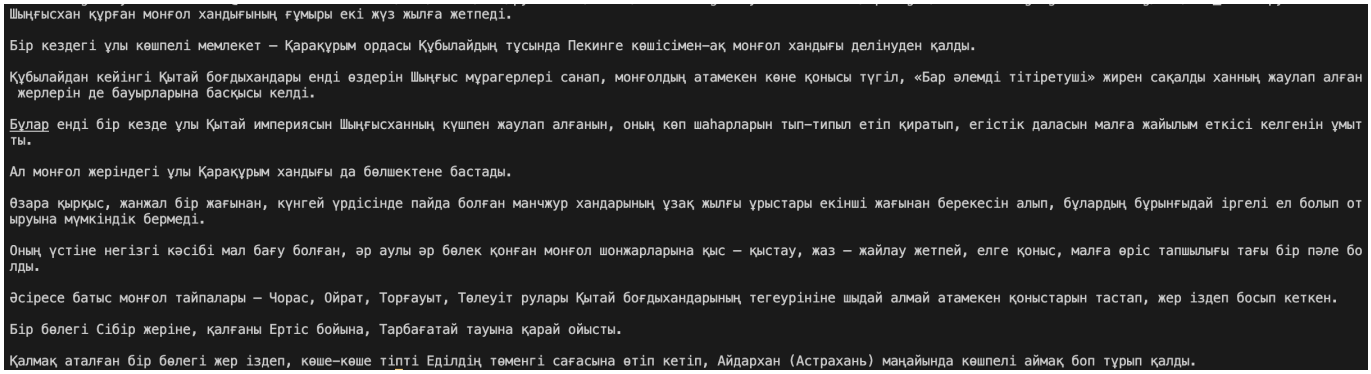

Figure 7: Text from 'Көшпенділер: Жанталас'.

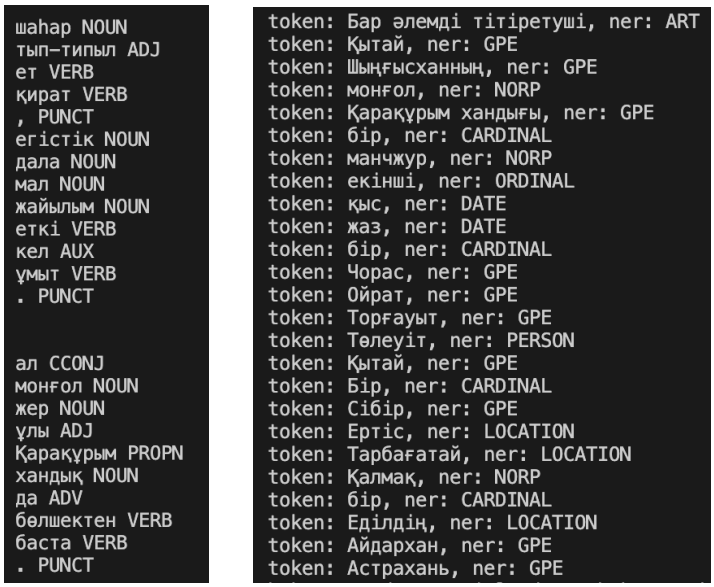

Figure 8: Lemmatization, POS tagging, and NER results

Stanza was able to evaluate 'Қарақұрым хандығы' as a geopolitical entity. It also could distinguish between a geopolitical entity and a location. Entities in the text date back to XVII-XVIII centuries.