



Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus

Kahaerjiang Abiderexiti, Maihemuti Maimaiti,
Tuergen Yibulayin, Aishan Wumaier

*School of Information Science and Engineering, Xinjiang
University, Urumqi, Xinjiang, 830046, China*

*Xinjiang Laboratory of Multi-Language information Technology,
Urumqi, Xinjiang, 830046, China*



OUTLINE

1. Introduction
2. UyNe Annotation Schemes
3. UyNeRel Annotation Schemes
4. Corpus Sampling and Annotation
5. Conclusions



1.Introduction

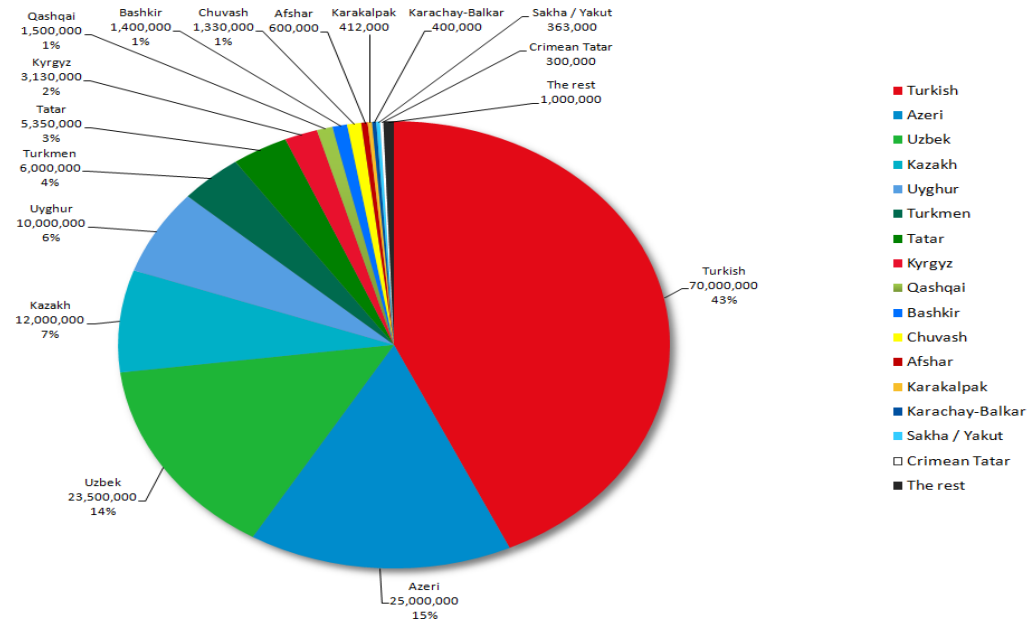
Uyghur language :

➤ is one of the official language as well as Mandarin (Chinese).

➤ fifths big language among Turkic languages



Number of Native Speakers in the Turkic Language Family





1.Introduction

Uyghur language :

➤ is one of the official language as well as Mandarin (Chinese).

➤ fifths big language among Turkic languages





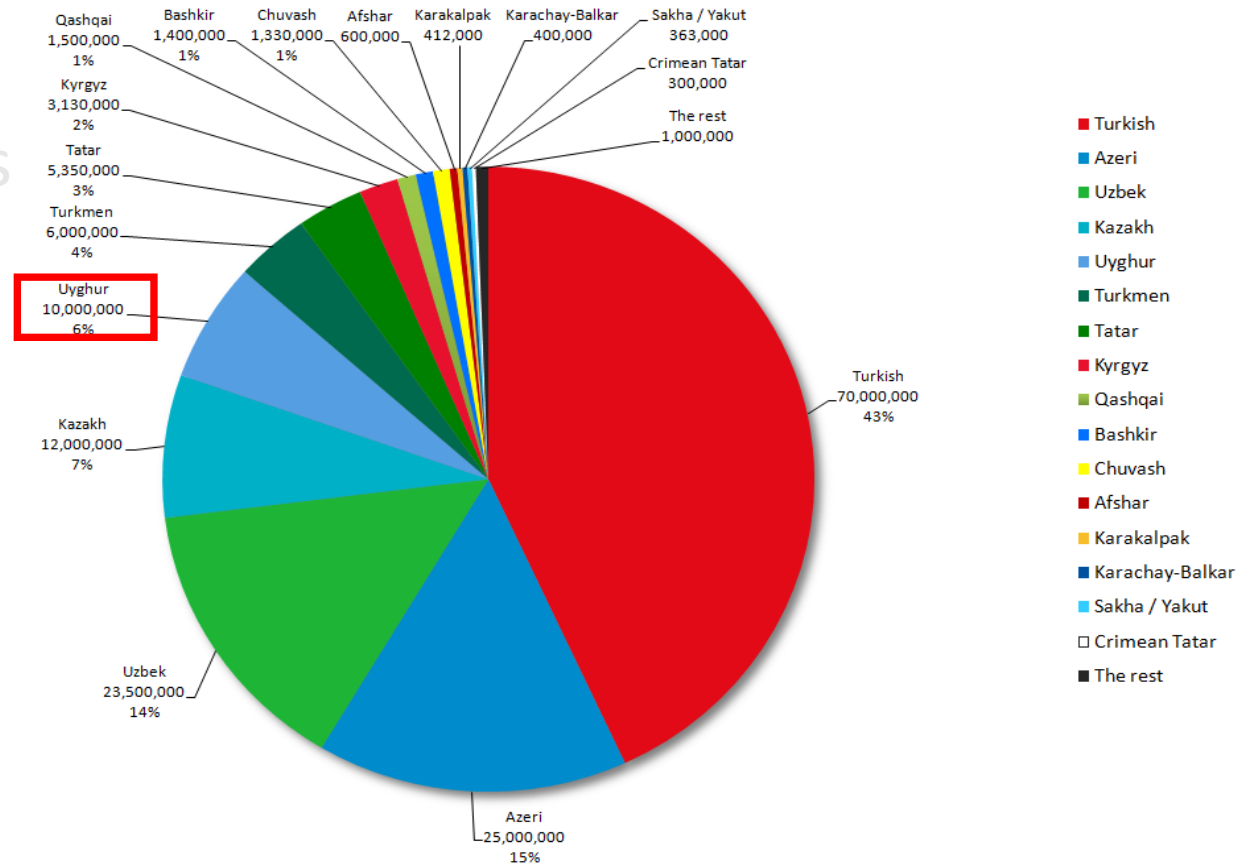
1.Introduction

Uyghur language :

➤ is one of the official language as well as Mandarin (Chinese).

➤ fifths big language among Turkic languages

Number of Native Speakers in the Turkic Language Family



http://en.wikipedia.org/wiki/Turkic_languages



1.Introduction

- **Entity**

An object or set of objects in the world.

- Entity Type:

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE),location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation:

relations of the targeted types between entities.

E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type:

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation, China, Tian Shan mountain.

- Entity Relation:

relations of the targeted types between entities.

E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation, China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:

Scenery at Altun Mountains Nature Reserve in Xinjiang





1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:

Scenery at  Altun Mountains Nature Reserve ⁱⁿ  Xinjiang



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:

Scenery at **Altun Mountains Nature Reserve** in **Xinjiang**

The diagram shows two orange boxes with arrows pointing up to the words 'Reserve' and 'Xinjiang' respectively. The left box is labeled 'LOC' and the right box is labeled 'GPE'.



1.Introduction

- Entity

An object or set of objects in the world.

- Entity Type

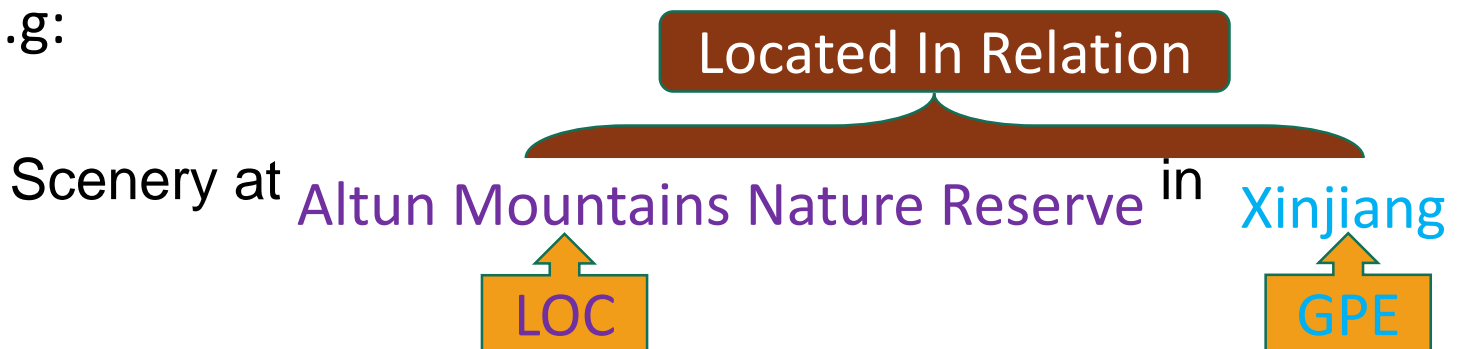
Person name(PER), Organization name(ORG), geographical/geopolitical entity name(GPE), location(LOC) etc.

E.g. Trump, United Nation ,China, Tian Shan mountain.

- Entity Relation

relations of the targeted types between entities.

E.g:





1.Introduction

To the best of our knowledge, currently

No Uyghur Named-Entity Relation Corpora



2. Uyghur Named Entity Annotation (UyNe) Scheme

1) Types of Entities

We annotate entities of Person(**PER**), Organization(**ORG**), GeoPolitical Entities(**GPE**), Location(**LOC**) and Facility(**FAC**).



2. Uyghur Named Entity Annotation (UyNe) Scheme

1) Types of Entities

We annotate entities of Person(**PER**), Organization(**ORG**), GeoPolitical Entities(**GPE**), Location(**LOC**) and Facility(**FAC**).

Ex:

تارىم دەرياسى
Tarim River

فرانسىيە
France

ئابدۇرېھىم ئۆتكۈر
Abdurehim Otkur

گۇگۇڭ سارىيى
Forbidden City

خۇاۋېي شىركىتى
Huawei Corporation



2. Uyghur Named Entity Annotation (UyNe) Scheme

2) Basic Annotation Unit

Surface form as the basic entity annotation unit.



2. Uyghur Named Entity Annotation (UyNe) Scheme

2) Basic Annotation Unit

Surface form as the basic entity annotation unit.

Ex:

سەپپىدىن ئەزىزنىڭ ئەسلىمىسى

↑
PER

Saypidin Azizi 's memoirs

↑
PER



2. Uyghur Named Entity Annotation (UyNe) Scheme

3) UyNe Annotation Rules

Such as entities are annotated according to usage.

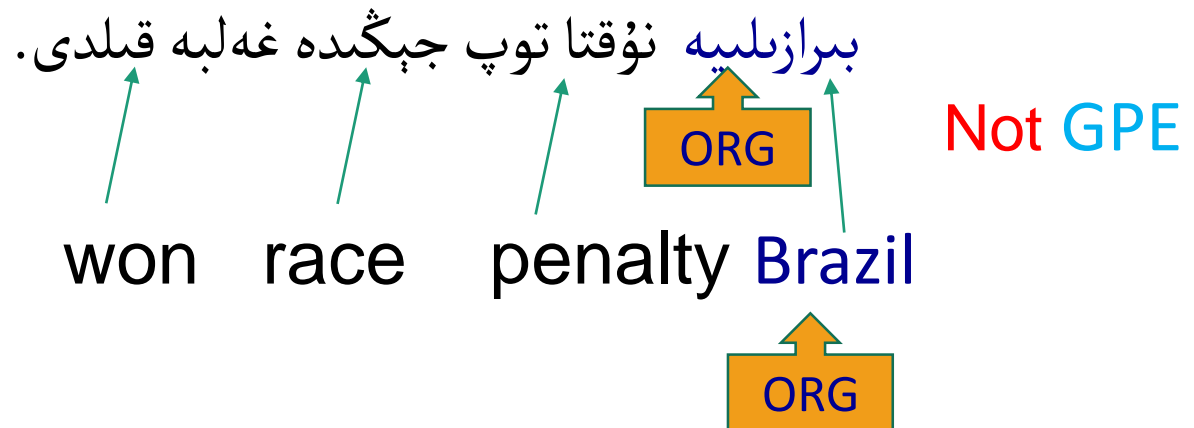


2.Uyghur Named Entity Annotation (UyNe) Scheme

3) UyNe Annotation Rules

Such as entities are annotated according to usage.

Ex:





3. Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

1) Types of UyNeRel

In UyNeRel there are **5 types** of relations. These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation). There is **15 subtypes**.



3. Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

1) Types of UyNeRel

In UyNeRel there are **5 types** of relations. These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation).

There are **15 subtypes** .

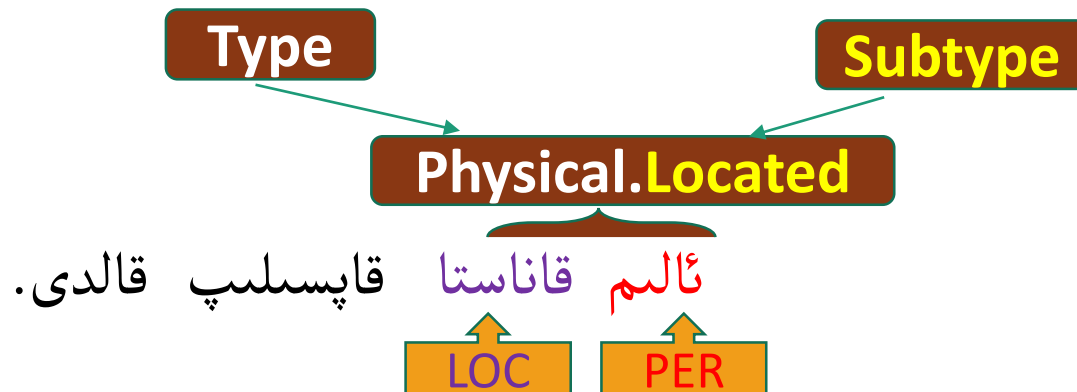


3.Uyghur Named Entity Relation (UyNeRel) Annotation Scheme

1) Types of UyNeRel

In UyNeRel there are **5 types** of relations. These are Physical, PartWhole, Gen-Aff (General-Affiliation), Per-social, Org- Aff (Organization-Affiliation).

There are **15 subtypes** . Ex:



(Alim was trapped in Kanas.)



4. Corpus Sampling and Annotation

Crawl raw corpus from websites

- <http://uy.ts.cn>
- <http://uyghur.people.com.cn>
- <http://uyghur.news.cn>

خەلق تورى ۱۸ - نۆۋەتلىك مەركىزىي كومىتېتىنىڭ 5 - ئومۇمىي يىغىنىغا نەزەر

10 - ئاينىڭ 26 - كۈنىدىن 29 - كۈنىگىچە ئېچىلىدۇ

خەلق تورى | مۇھىم مەملىكەت | شىنجاڭ | خەلقئارا | سىن | مۇزىتەلەر | ئوبزور | خەۋىرىمىز | مائارىپ | دىن | پازۇغۇچىلىرىمىز

ئەدەبىيات | كىتەپ | كىچ | مەدەنىيەت | شەخسلەر | ساياھەت | يېمەك - ئىچمەك | كۆڭۈل چىشىش | ئۆزۈمۈش | سۆز قاندىرۇش

ئۆزلىك خەۋەرلەر | بەشى پىرۋىتلاش پىلانلاش جۇڭگوچە شەھەر تەرەققىيات يولىغا بېتەكچىلىك قىلىشى | «بوزكۆت»

خەلق تورىنىڭ ئىمىلى Email ئادرېسى: ئىختىيارى مۇخبىرلارغا خەت

شى جىنپىڭ خەت يېزىپ جۇڭگو تېبابىتى پەنلەر ئاكادېمىيەسى قۇرۇلغانلىقىنىڭ 60 يىللىقىنى تەبرىكلەيدى

• مەركەزلىك شەھەر خىزمىتى بىغىنى بېيجىڭدا ئۆت...
 • بېيجىڭ شەھىرىنىڭ ئورگان كەسپى ئورۇنلىرى ئى...
 • مەملىكەتلىك خەلق قۇرۇلتىيى ئىسمى كومىتېتى...
 • مەركىزىي مىللەتلەر ئۇنىۋېرسىتېتى مۇزىكا ئىن...
 • مائارىپ مىنىستىرلىقى ئالاھىدە تۈرگە ئوقۇتقۇچى...
 • ئۈرۈمچى سانجى تەرەپ بىر ئاپتونۇس پانچۇقى...
 • ئۈرۈمچى بىلەن ئوسكا ئىككى جايلىك ساياھەت كە...
 • شىنجاڭ مۇسۇلمانلىرى «ۋىزىنىلىك ئۈنچى سانجى...
 • شەھەرلەرنى ئىشقا ياشاشقا مەن كېلىدىغان ھەم...
 • جىننەنكى 10 ياشلىق قىز دۇنيادىكى ئەڭ كىچىك...

ئۇيغۇر ۋاقتىدا English Uyghurche 中文

ئۇيغۇر ۋاقتىدا ۋەزىر دۇنياغا ئەزىز كۆزۈل شىنجاڭ ئالتاغلار مەدەنىيەت قاتنى مەخسۇس ساھىبە

خەۋەرگەر شىنجاڭ مەملىكەت خەتتارا كۆزۈل شىنجاڭ ساياھەت سەلىكە سۈرەتتىكى شىنجاڭ دۇنياغا ئەزىز سۈرەتتىكى دۇنيا مودا ۋە ئېسىم قارايمىنلار كەمبەيەت ئەدەبىي نەسىر ئىشور تەسىرەر تۇرمۇش خەتتارا ئۇمىق ئۇمىق تۇرمۇش ئۇمىق ئۇمىق قاتنى ئىزاد ئۇچۇرلار سايلانلار

دۇنيا ئالاقە تورى يىغىنى | 2015 - 2 - نۆۋەتلىك | 16 - ئېتىراپتىن 18 - ئېتىراپتىنچە جىيەت ۋەجىن

شىنجاڭ ئۇيغۇر تىلىنى قولغا ئېلىش ۋە يېڭىلىش ئۇيغۇر تىلىنى قولغا ئېلىش ۋە يېڭىلىش ئۇيغۇر تىلىنى قولغا ئېلىش ۋە يېڭىلىش

شەھەرلەرنى ياشاشقا تېخىمۇ ماس كېلىدىغان ، تېخىمۇ گۈزەل قىلىپ قۇرۇپ چىقىش كېرەك

ئەنئەنىۋىي ئاساسىي شەھەر رايونىدا بىر تەبىئىي گاز ئۆزۈمۈشى يېرىلىپ كەتتى

بىش خەۋەر

شى جىنپىڭ جۇڭگو جۇڭگو تېبابىتى پەنلەر ئاكادېمىيەسى قۇرۇلغانلىقىنى تەبرىكلەيدى

• مەركەزلىك شەھەر خىزمىتى بىغىنى بېيجىڭدا ئۆت...
 • مەملىكەتلىك خەلق قۇرۇلتىيى ئىسمى كومىتېتى...
 • مەركىزىي مىللەتلەر ئۇنىۋېرسىتېتى مۇزىكا ئىن...
 • مائارىپ مىنىستىرلىقى ئالاھىدە تۈرگە ئوقۇتقۇچى...
 • ئۈرۈمچى سانجى تەرەپ بىر ئاپتونۇس پانچۇقى...
 • ئۈرۈمچى بىلەن ئوسكا ئىككى جايلىك ساياھەت كە...
 • شىنجاڭ مۇسۇلمانلىرى «ۋىزىنىلىك ئۈنچى سانجى...
 • شەھەرلەرنى ئىشقا ياشاشقا مەن كېلىدىغان ھەم...
 • جىننەنكى 10 ياشلىق قىز دۇنيادىكى ئەڭ كىچىك...



4. Corpus Sampling and Annotation

Annotation Tool

UyNeRelAnnotation : 1.0V E:\200CON.TXT

OPEN SAVE

فرانسىيە زوڭتۇڭ مەھكىمىسى 1-فېۋرال مۇنۇلارنى جاكارلىدى. فرانسىيە زوڭتۇڭى
 هوللاندى ، گېرمانىيە زوڭلىسى مېرىكل ۋە ئۇكرائىنا زوڭتۇڭى پروشىنكو شۇ كۈنى 45 مىنۇت
 تېلېفونلاشتى.
 ئۈچ رەھبەر ئۇكرائىنانىڭ شەرقىدىكى توقۇنۇشنى تىزىدىن توختىتىشنى مۇراجەت قىلدى.
 ئۇكرائىنانىڭ سابىق زوڭتۇڭى كوچىما 31-يانۋار بىلورۇسىيەنىڭ پايتەختى مىسكىدا مۇنداق دېدى:
 شۇ كۈنى ئۆتكۈزۈلگەن ئۇكرائىنا مەسلىسى بويىچە ئۈچ تەرەپ بىرلەشمە گۇرۇپپىسى سۆھبىتى
 ئۇكرائىنانىڭ شەرقىدىكى ئىل ئىچى قۇراللىق تەشكىلاتى سەۋەبىدىن بۇزۇلدى.
 هوللاندى ، مېرىكل ۋە پروشىنكو ئۇكرائىنا ۋەزىيىتىنى مۇزاكىرە قىلىپ ، ئۇكرائىنا مەسلىسى بويىچە

Entity

Delete 28~19|||ئۇكرائىنا GPE

Comment

Delete 550~528|||ئۇكرائىنانىڭ شەرقىدىكى LOC

Comment

Relation

PartWhole Geo Insert Relation

ID	EType	StartPos	EndPos	EText	MentionLevel	ETextID	EFileName	Comment	ETime	IsDel
2016...	GPE	0	8	فرانسىيە	NAM	2016...	200CON		2016...	<input type="checkbox"/>
2016...	GPE	9	18	گېرمانىيە	NAM	2016...	200CON		2016...	<input type="checkbox"/>
2016...	GPE	19	28	ئۇكرائىنا	NAM	2016...	200CON		2016...	<input type="checkbox"/>
2016...	LOC	40	62	ئۇكرائىنانىڭ شەرقىدىكى	NAM	2016...	200CON		2016...	<input type="checkbox"/>

ID	FromID	FromText	ToID	ToText	RType	RSubType	RTime	RTextID	RFileName	Comment	IsDel
2016...	201...	پروشىنكو	2016...	زوڭتۇڭى	PerSocial	Role	2016...	2016...	200CON		<input type="checkbox"/>
2016...	201...	هوللاندى	2016...	فرانسىيە	OrgAff	Employment	2016...	2016...	200CON		<input type="checkbox"/>
2016...	201...	گېرمانىيە	2016...	مېرىكل	OrgAff	Employment	2016...	2016...	200CON		<input type="checkbox"/>
2016...	201...	پروشىنكو	2016...	ئۇكرائىنا	OrgAff	Employment	2016...	2016...	200CON		<input type="checkbox"/>



4. Corpus Sampling and Annotation

Annotated XML text Format:

ا مەسىلىسى بويىچە ئۈچ تەرەپ بېلورۇسىيىنىڭ پايتەختى مىنىسكىدا مۇنداق دېدى: شۇ ئۇكرائىنانىڭ سابىق زوڭتۇڭى كوچىما 31- يانۋار
سكى سۆھبىتىنىڭ مەغلۇب مۇزاكىرە قىلىپ، ئۇكرائىنا مەسىلىسى بويىچە ئۈچ تەرەپ ھوللاندى، مېرىكىل ۋە پروشىنكو ئۇكرائىنا ۋە زېپىتىنى
]]></TEXT>

<TAGS>

```

<GPE id="GPE0" spans="723~730" text="مىنىسكى" mentionLevel="NAM" comment="" />
<GPE id="GPE1" spans="722~731" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
<GPE id="GPE2" spans="665~674" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
<GPE id="GPE3" spans="628~637" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
<PER id="PER0" spans="619~627" text="پروشىنكو" mentionLevel="NAM" comment="" />
<PER id="PER1" spans="608~615" text="مېرىكىل" mentionLevel="NAM" comment="" />
<PER id="PER2" spans="598~606" text="ھوللاندى" mentionLevel="NAM" comment="" />
<GPE id="GPE4" spans="559~568" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
<ORG id="ORG0" spans="551~578" text="ئەل ئىچى قۇراللىق تەشكىلاتى" mentionLevel="NAM" comment="" />
<LOC id="LOC0" spans="528~550" text="ئۇكرائىنانىڭ شەرقىدىكى" mentionLevel="NAM" comment="" />
<GPE id="GPE5" spans="528~540" text="ئۇكرائىنانىڭ" mentionLevel="NAM" comment="" />
<GPE id="GPE6" spans="465~474" text="ئۇكرائىنا" mentionLevel="NAM" comment="" />
<GPE id="GPE7" spans="422~431" text="مىنىسكىدا" mentionLevel="NAM" comment="" />
<GPE id="GPE8" spans="399~412" text="بېلورۇسىيىنىڭ" mentionLevel="NAM" comment="" />
<PER id="PER3" spans="381~387" text="كوچىما" mentionLevel="NAM" comment="" />
<TTL id="TTL0" spans="373~380" text="زوڭتۇڭى" comment="" />
<TTL id="TTL1" spans="367~380" text="سابىق زوڭتۇڭى" comment="" />
<GPE id="GPE9" spans="354~366" text="ئۇكرائىنانىڭ" mentionLevel="NAM" comment="" />

```



4. Corpus Sampling and Annotation

Cohen's K inter-annotator agreement (IAA):

Preliminary result:

$$\kappa_{UyNe} = 0.7$$

$$\kappa_{UyNeRel} = 0.6$$

Now it is higher than 0.8.

500 documents are annotated.

The data available :

<https://github.com/kaharjan/UyNeRel>



6. Conclusion

- Investigate theoretical foundations, sample raw text from internet sources for Uyghur language
- Developed Uyghur UyNeRel annotation schemes and tool
- The annotated corpus is publicly available



Thank You !