

# MONGOLIAN TAGSET and CORPUS TAGGING

J.Purev and Ch. Odbayar

**CRLP**

Center for Research on Language Processing  
National University of Mongolia

**(NUM)**

# OUTLINE

- Introduction to NUMTeam project
- POS Tagset for Mongolian
- Mongolian Corpus Tagging
- Conclusion

# NUM Team

- NUMTeam, Mongolia
  - At Center for Research on Language Processing (CRLP)
    - First center in Mongolia
    - Established in May 2007
    - Supported by National University of Mongolia and PANLocalization
    - Currently 7 researchers and staffs including computer scientists and linguists
    - Goal: *Mongolian Localization and Processing*
  - Our team joint in PANL10n Project in 2006
- Project Objectives:
  - 5 million tagged words corpus
  - Cleaning Tools and Spell Checker
  - POS Tagset
  - POS Tagger

# MONGOLIAN LANGUAGE

- Mongolian is a highly agglutinative language
  - More than 200 inflectional suffixes
  - Case marking
  - Variations in inflectional forms
  - Vowel harmony, etc

# POS TAGSET DESIGN

- Follow PennTree Bank to design tag marks
- Two-level tagset
  - High level
  - Low level
- High level is similar to English
  - N (noun), морь (horse)
  - V (verb), унах (ride)
  - JJ (adjective), etc
- Low level is
  - G (genitive), -ийн/ -ний/ -ын/ -ны/ -ий/ -ы
  - D (past tense), -сан/ -лаа/ -в
  - S (possessive), etc
- Using high and low
  - NG (noun genitive), морины (of horse)
  - VD (verb past), унасан (rode)
  - NGS (noun genitive + possessive), морныхоо (of my horse)

# HIGH-LEVEL TAGSET

- Totally 23 tags

High Level POS Tagset								
Noun			Verb			Other		
1.	Noun	N	12.	Verb	V	19.	Co-conjunction	CC
2.	Pronoun	PN	13.	Proverb	PV	20.	Sub-conjunction	CS
3.	Proper noun	RN	14.	Adverb	RB	21.	Interjection	INTJ
4.	Adjective	JJ	15.	Pro-adverb	PRB	22.	Question	QN
5.	Pro-adjective	PJ	16.	Ad-adverb	RBA	23.	Punctuation	PUN
6.	Ad-adjective	JJA	17.	Modal	MD			
7.	Superlative	JJS	18.	Auxiliary	AUX			
8.	Number	C						
9.	Preposition	PR						
10.	Postposition	PT						
11.	Abbreviation	ABR						

# LOW-LEVEL TAGSET

- Following tags are combined with the high-level tags and each other
- Totally 31 tags

Low Level POS Tagset					
Noun			Verb		
1.	Nominative	N	16.	Active	A
2.	Genitive	G	17.	Passive	P
3.	Locative	L	18.	Plural	P
4.	Accusative	C	19.	Aspect	A
5.	Ablative	B	20.	Past	D
6.	Instrumental	I	21.	Present	P
7.	Commutative	M	22.	Continues	G
8.	Single	/	23.	Future	F
9.	Plural	P	24.	Infinitive/Base	B
10.	Count	/	25.	Coordination	C
11.	Uncount	U	26.	Subordination	S
12.	Possessive	S	27.	1	1
13.	Comparative	R	28.	2	2
14.	Approximate	A	29.	3	3
15.	Abbreviated noun possessive	H	30.	Negative	X
			31.	Comparative	R

# USING HIGH and LOW

- Currently, around 130 combination tags are created
  - Most of them are tags for noun and verb inflections
  - Tag marking length is 1 - 5
- Some examples of the tags created while tagging the corpus

Tag	Meaning	Mongolian	English
N	Noun	морь	horse
NB	Noun Ablative	мориноос	from horse
NBS	Noun Ablative Possessive	мориноосоо	from my horse
NC	Noun Accusative	морийг	horse (accusative)
NCS	Noun Accusative Possessive	морийгоо	my horse (accusative)
ND	Noun Direction	талруу	to field
NDS	Noun Direction Possessive	талруугаа	to my field
NG	Noun Genitive	морний	of horse
NGHB	Noun Genitive Special-possessive Ablative	орныхоос	from someone's country
NGHMS	Noun Genitive Special-possessive Commutative Possessive	орныхтойгоо	with own country's something
NGS	Noun Genitive Possessive	морнийхоо	of my horse

# USING HIGH and LOW

- Currently, around 130 combination tags are created
  - Most of them are tags for noun and verb inflections
- Some examples of the tags created while tagging the corpus

Tag	Meaning	Mongolian	English
N	Noun	морь	horse
NB	Noun Ablative	мориноос	from horse
NBS	Noun Ablative Possessive	мориноосоо	from my horse
NC	Noun Accusative	морийг	horse (accusative)
NCS	Noun Accusative Possessive	морийгоо	my horse (accusative)
ND	Noun Direction	талруу	to field
NDS	Noun Direction Possessive	талруугаа	to my field
NG	Noun Genitive	морний	of horse
NGHB	Noun Genitive Special-possessive Ablative	орныхоос	from someone's country
NGHMS	Noun Genitive Special-possessive Commutative Possessive	орныхтойгоо	with own country's something
NGS	Noun Genitive Possessive	морнийхоо	of my horse

# USING HIGH and LOW

- Currently, around 130 combination tags are created
  - Most of them are tags for noun and verb inflections
- Some examples of the tags created while tagging the corpus

Tag	Meaning	Mongolian	English
N	Noun	морь	horse
NB	Noun Ablative	мориноос	from horse
NBS	Noun Ablative Possessive	мориноосоо	from my horse
NC	Noun Accusative	морийг	horse (accusative)
NCS	Noun Accusative Possessive	морийгоо	my horse (accusative)
ND	Noun Direction	талруу	to field
NDS	Noun Direction Possessive	талруугаа	to my field
NG	Noun Genitive	морины	of horse
NGHB	Noun Genitive Special-possessive Ablative	орныхоос	from someone's country
NGHMS	Noun Genitive Special-possessive Commutative Possessive	орныхтойгоо	with own country's something
NGS	Noun Genitive Possessive	мориныхоо	of my horse

# MAIN CHARACTERISTICS

- Gerund is inflected all noun inflectional suffixes

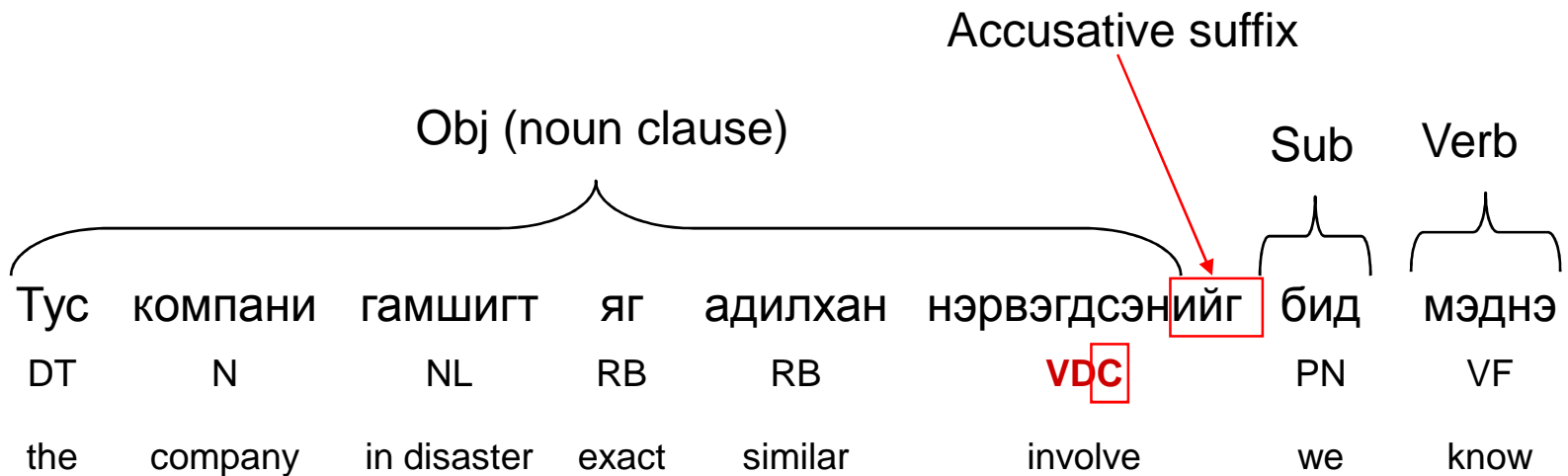
This sentence consists of noun and main clauses  
Noun clause's verb is inflected with accusative case to function as a object of the main clause  
So the main verb INVOLVE is not tagged just as past tense verb – VD instead it should be tagged VDC, past tense verb plus accusative case

Тус	компани	гамшигт	яг	адилхан	нэрвэгдсэнийг	бид	мэднэ
DT	N	NL	RB	RB	<b>VDC</b>	PN	VF
the	company	in disaster	exact	similar	involve	we	know

We know the company was involved in the disaster similarly

# MAIN CHARACTERISTICS

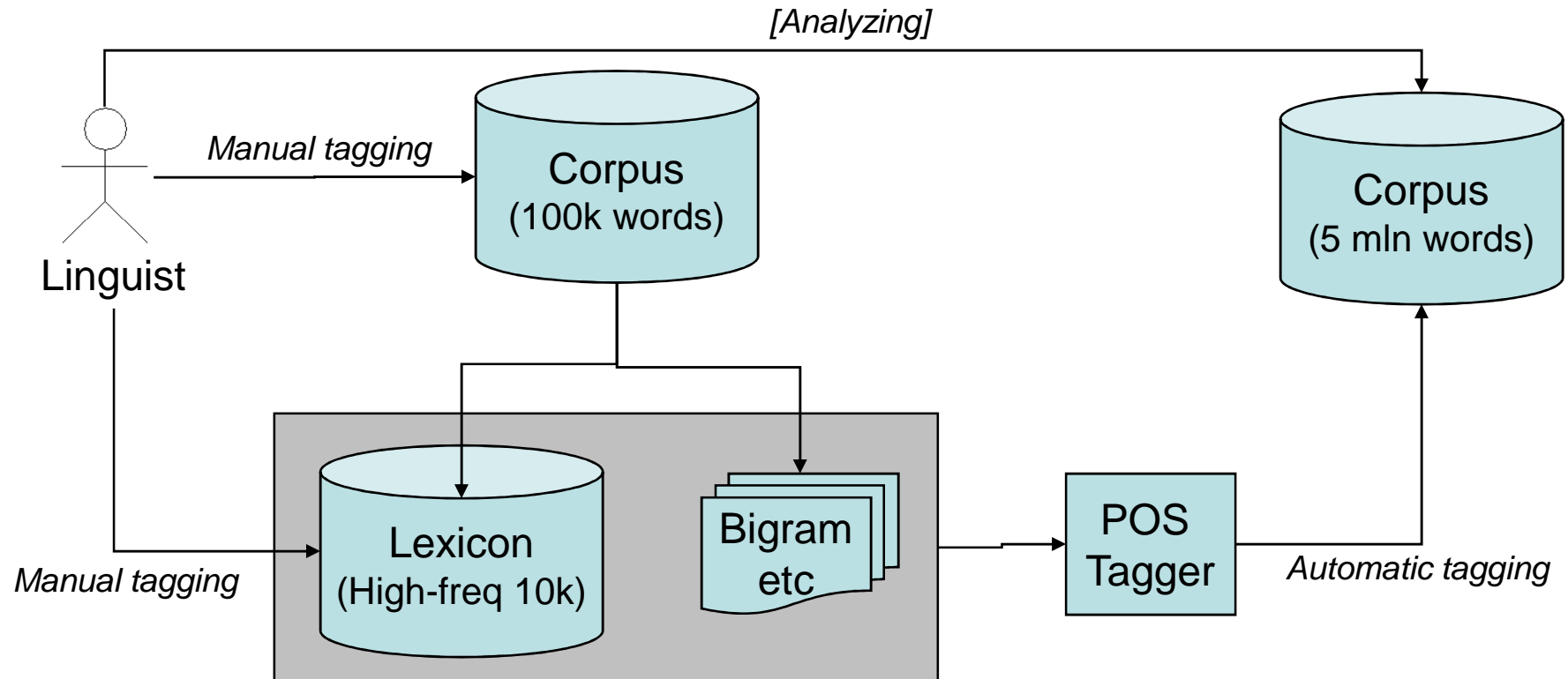
- Gerund is inflected all noun inflectional suffixes



We know the company was involved in the disaster similarly

# CORPUS TAGGING

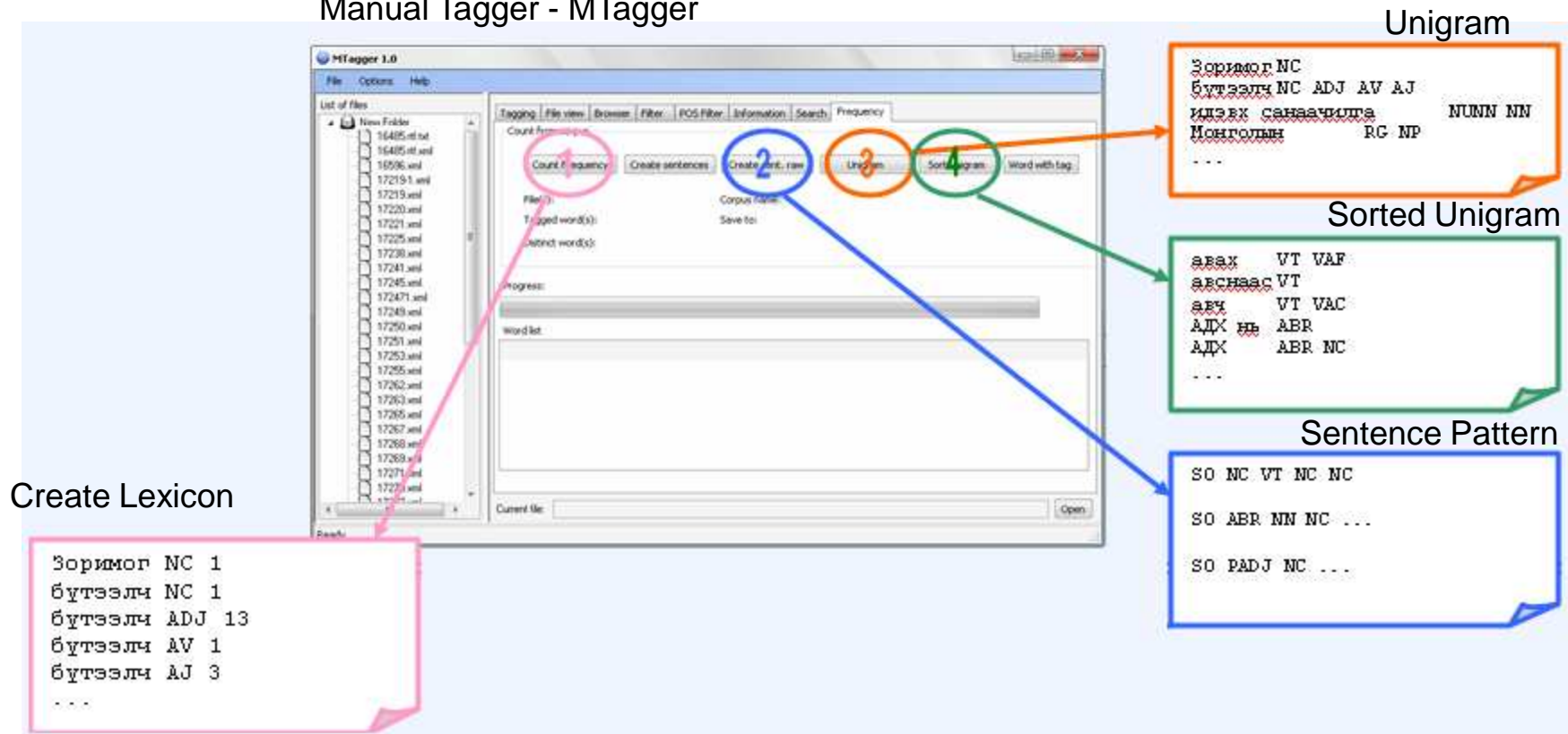
First, the linguist manually tag some part of the corpus  
And training the POS tagger with that manually tagged part  
And the whole corpus is automatically tagged  
Lastly, linguist analyzes the automatically tagged corpus



# TRAINING DATA [1]

- To prepare the training data /unigram, sentence pattern, lexicon/ for the tagger
- MTagger (manual tagger) is used

Manual Tagger - MTagger



```
Зоримог NC 1
бүтээлч NC 1
бүтээлч ADJ 13
бүтээлч AV 1
бүтээлч AJ 3
...
```

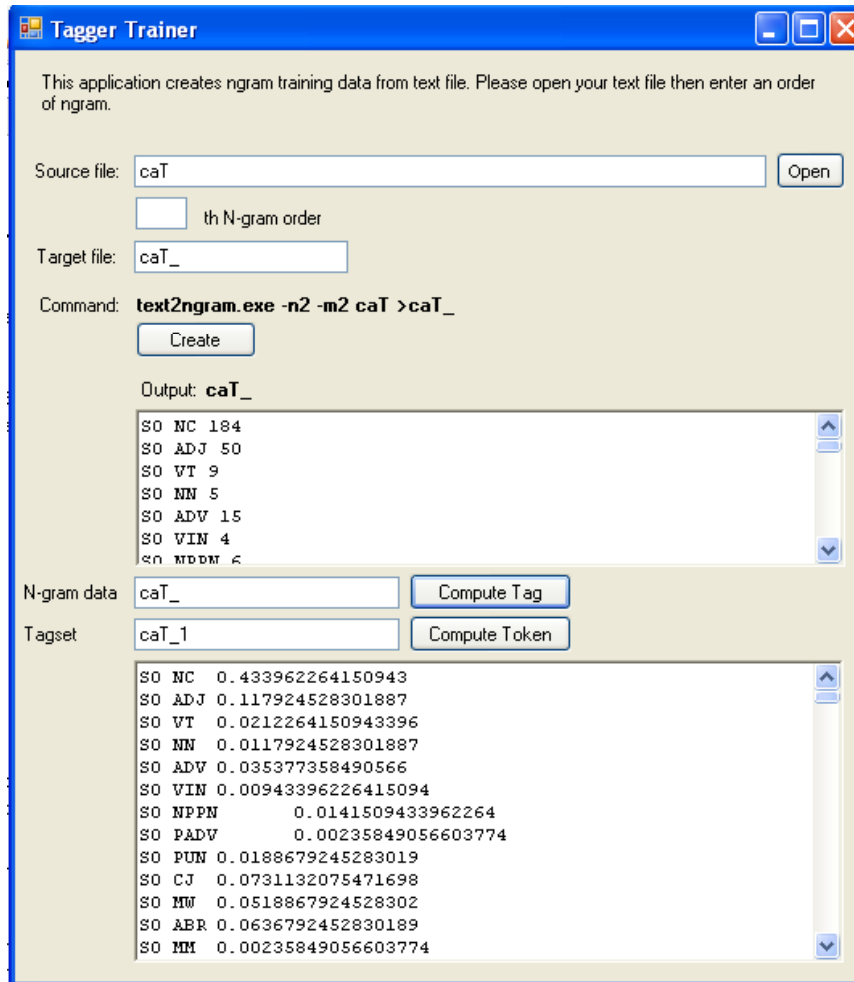
```
Зоримог NC
бүтээлч NC ADJ AV AJ
ИДЭХ санаачидга NUNN NN
Монголын RG NP
...
```

```
саях VT VAF
сечнаас VT
саях VT VAC
АДХ нь ABR
АДХ ABR NC
...
```

```
SO NC VT NC NC
SO ABR NN NC ...
SO PADJ NC ...
```

# TRAINING DATA [2]

After producing the training data  
Bigram and lexicon is produced for POS Tagger



Training Data used in POS Tagger

SO NC 179  
SO RC 4  
SO AJ 22  
...  
NC ADV 317  
NC ADJ 497  
...

SO 534  
NC 5502  
NUNN 1  
RG 20  
NPL 14  
...

SO NC 0.33520599250936  
SO RC 0.00749063670411  
...

Зорилгог NC 0.0001817  
Бүтээлч NC 0.0001817  
Бүтээлч ADJ 0.0090403  
...

# CONCLUSION

- In this phase
  - Designed POS Tagset for Mongolian
    - Two level tagset: High-level and Low-level
      - High-level 23 tags such as Noun, Verb, Adjective, etc
      - Low-level 31 tags such as Genitive, Tenses, etc
      - Currently, actual 130 tags and it increases further
    - Main feature is tag for inflectional suffixes' marks
  - Tagged the corpus
    - Manually tagged around 150 thousand words
    - Developed tools for training data (bigram)
    - Automatically tagged the corpus by Bigram POS Tagger
    - Manual checking the automatically tagged corpus is ongoing at this time

Thank you