



An Introduction to NLP research in PANL10n Phase II

Purev Jaimai

**Professor and Head
Computer and IT Department
School of Information Technology**

[SIT]

National University of Mongolia

[NUM]

purev@num.edu.mn

Structure

- Background
National University of Mongolia
School of IT
Mongolian Language
- NUM Team
- NLP Research Work Organization (**What How Do**)
- Some problem, challenges

National University of Mongolia

- Was founded in 1942
- Amount of students: ~10 thousand
 - Master students: **1030**
 - PhD students: **431**
 - International students: **222 Undergraduates / Graduates / Doctorates**

(Russia, China, Japan, Korea, USA, Poland, Vietnam, Czech Republic ...)

- Faculties/Schools: **12**
- Staffs: **1058**
 - **650** lecturers
 - **250** staffs

International Affairs

- **Number of relationships:** 129 universities/organizations
 - Asia: 65
 - America: 8
 - Europe: 56
- **International Networks:**
 - International Association of Universities
 - University Mobility in Asia and Pacific

 - Euro-Asia Pacific University Network
 - Euro-Asia University Network
 - Council on International Educational exchange
 - Union of Northeast Asian 5 Universities
 - Asia University Federation

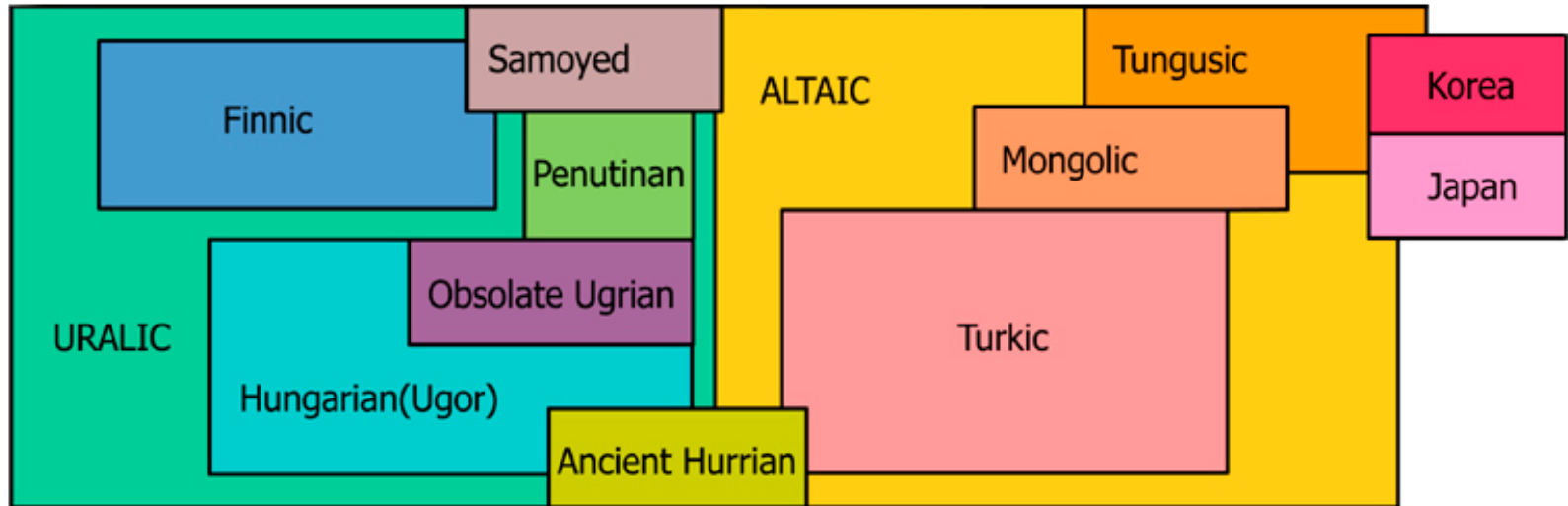
School of Information Technology

- First establishment was in 1990 as Department of Electronics at School of Physics and Electronics
- SIT was established in 2002
- Departments:
 - Computer and Information Technology Department
 - Department of Electronics
- Staffs: around 30
- Amount of students: ~500



ABOUT MONGOLIAN LANGUAGE

Ural-Altaic languages



- the Mongolic language group has about 10 languages
- Most closely related to the Tungus and Turkish

Ural-Altaic languages /2

- Approximately 5 million people speak:
 - in Mongolia (2.5mln)
 - in China (1.5mln)
 - in Afghanistan (?)
 - in Russia (0.5mln)



Main Characteristics

- Sentence order: Subject-Object-Verb **SOV**
- No genders
- No gender for personal pronouns
- Vowel harmony /Word stem doesn't change/
- Agglutination

- Adjectives precede nouns
- There is hardly any difference between nouns and adjectives.
- 7 pre-nominal elements can be placed in front of noun

Main Characteristics /2

- Suffixes are very productive

Suffixes inflectional and derivational are joined with all the possible versions from the actively used verb “cyp” (to study, to learn). There were estimating that number reached to 550.

- Each verb in Mongolian can be converted more than 150 forms
- Maximum, 11 suffixes can be added to one root
- Changing only vowels, then new word is able to be created.
 - Давжаа-Davjaa (small, little)
 - Довжоо-Dovjoo (porch)
 - Дэвжээ-Devjee (ring)

Mongolian Phonemes

- 47 phonemes
 - 14 of them are vowels
 - 7 short vowels
 - 7 long vowels
 - 33 consonants

Mongolian Phonemes /2

- SYLLABLES (C=consonant, V=vowel)
 - Genuine Mongolian words use the following syllable structure:
 - V - syllables
 - VC - syllables
 - CV - syllables
 - CVC - syllables

Mongolian Phonemes /3

- Through the *Cyrillic* script, the following artificial syllables came into usage:
 - VCC - syllables
 - CVCC - syllables
 - VCCC - syllables
 - CVCCC - syllables

Mongolian Word Segmentation

- Mongolian words are relatively easily detected from the text since a space is supposed to be placed between them.
- A sentence begins with capital letter.
- Full stop (.) marks the end of a sentence.

Mongolian Script

- the Mongols created/used at least 10 scripts:
- Phagspa Script (14th century),
- Tod Script (16th century),
- Soyombo Script (17th century),
- Vagindra Script (19th century).

ᠮᠤᠩᠭᠣᠯᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ
ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ

ᠪᠠᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ
ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ
ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ
ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ ᠨᠠᠭᠤᠯᠠᠭᠤᠯᠤᠰᠤ

Phagspa script

Mongolian Script /2

- Today, the Mongolian language uses two official scripts:
 - The (new) Cyrillic Mongolian Script (so-called shortly Cyrillic Script)
 - The (old) Mongolian Script

Mongolian Script/3

- The traditional (old) Mongolian has 56 characters
- There are 7 basic vowels, 27 consonants

Vowels		Consonants					
	а а		н н		ш ш		х х
	э е		нг		т т		к к
	и и		б б		д д		ц ц
	о о		п п		ч ч		з з
	у у		х х		ж ж		х х
	ө ө		г г		й й		ж ж
	ү ү		м м		р р		л л
	ээ		л л		в в		з з
			с с		ф ф		ч ч

Mongolian Script /4

- The (new) Cyrillic Mongolian Script
 - 35 letters (35 small, 35 capital):

Бб (b)	Зз (ds)	Пп (p)	Хх (h)
Вв (v)	Кк (k)	Рр (r)	Цц (ts)
Гг (g)	Лл (l)	Сс (s)	Чч (ch)
Дд (d)	Мм (m)	Тт (t)	Шш (sh)
Жж (j)	Нн (n)	Фф (f)	Щщ (shch)

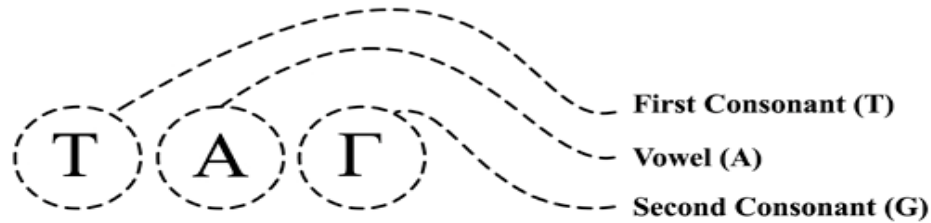
(a) Mongolian Consonants /20/

Аа (a)*	Ее (ye)	Ёё (yo)
Ии (i)*	Йй (y)	Оо (o)*
Өө (closed o)*	Уу (u)*	Үү (closed u)*
Ыы (ii)	Ээ (e)*	Юю (yu)
Яя (ya)		

(b) Mongolian Vowels /13/

ьь	ъъ
----	----

(c) Mongolian Signs /2/



(d) TAG in Mongolian

Seven basic vowels marked by asterisk (*).

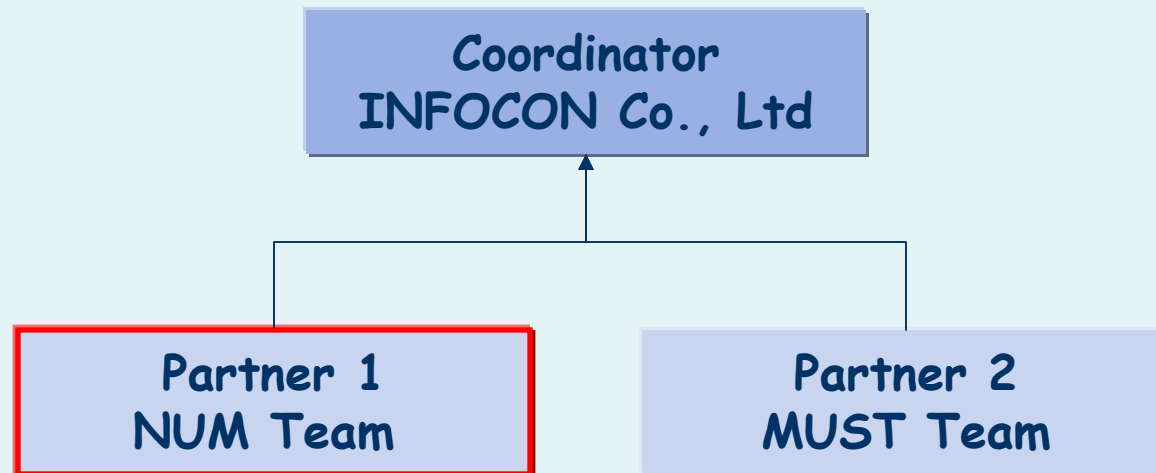


NUM TEAM

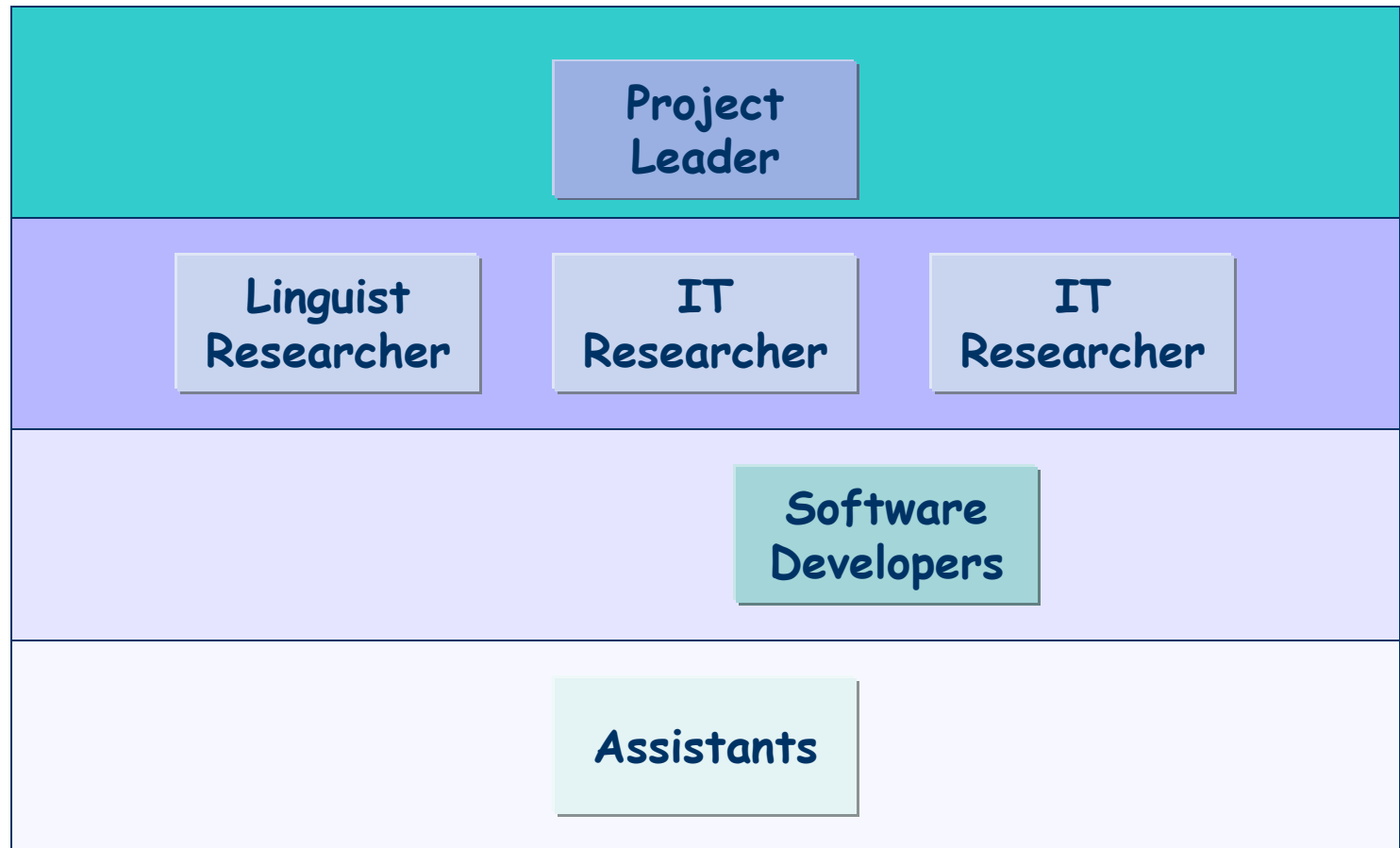
PANL10n Project II PHASE

Mongolian Teams Structure

COUNTRY COMPONENT - MONGOLIA



NUM Team Structure



Our Objective for Phase II

- Establishment of Language Processing Research Center
- Corpus with 5 Million Word
- 10k Word Lexicon
- Spell Checker
- POS Tagger
- Sentence Parser
- HR development

NLP RESEARCH WORK ORGANIZATION

[What How Do]

Project will be divided into 3 sub phases

II.1 Phase

- Deliverables:
 - 1 million word corpus
 - 100k word manually tagged corpus
- How to do:
 - Make corpus collection design
 - Collect texts according to the categories
 - Design tagset for Mongolian
 - Using and developing software

II.2 Phase

- Deliverables:
 - 5 million Word Corpus
 - 10k Word Lexicon
 - POS Tagger
 - Spell Checker
- How to do:
 - Add texts to Corpus Design
 - Analyze Frequencies of Words in Corpus
 - Use Tagset determined in Phase 1 for POS Tagger
 - Using and Developing Tools

II.3 Phase

- Deliverables:
 - Sentence Parser
 - Prototype Computational Grammar
- How to do:
 - Analyze the tagged corpus
 - Developing tools and software
 - Run the parser on the corpus
 - Analyze the parsed corpus

Expected Outputs of The Project

MONGOLIAN CORPUS DEVELOPMENT

SOFTWARE AND TOOLS FOR MONGOLIAN CORPUS

Corpus Organizer

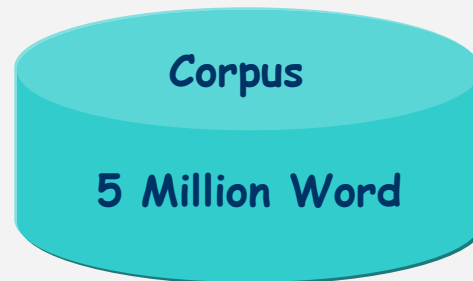
Corpus Analyzer

POS Tagger

Search Engine

Sentence Parser

Spell Checker



Some Problems

- NLP - new research and development area for Mongolia
- Not enough experience in NLP research and development.

Transfer of an experience, available source code and methods will be useful for saving time.

- Shortage of NLP trained human resource.
- No professional computational linguist.

It needs short and long term training on language technology



Thank you for attention!