Mongolian to Japanese machine translation using ChaSen

EHARA Terumasa*

HAYATA Suzushi

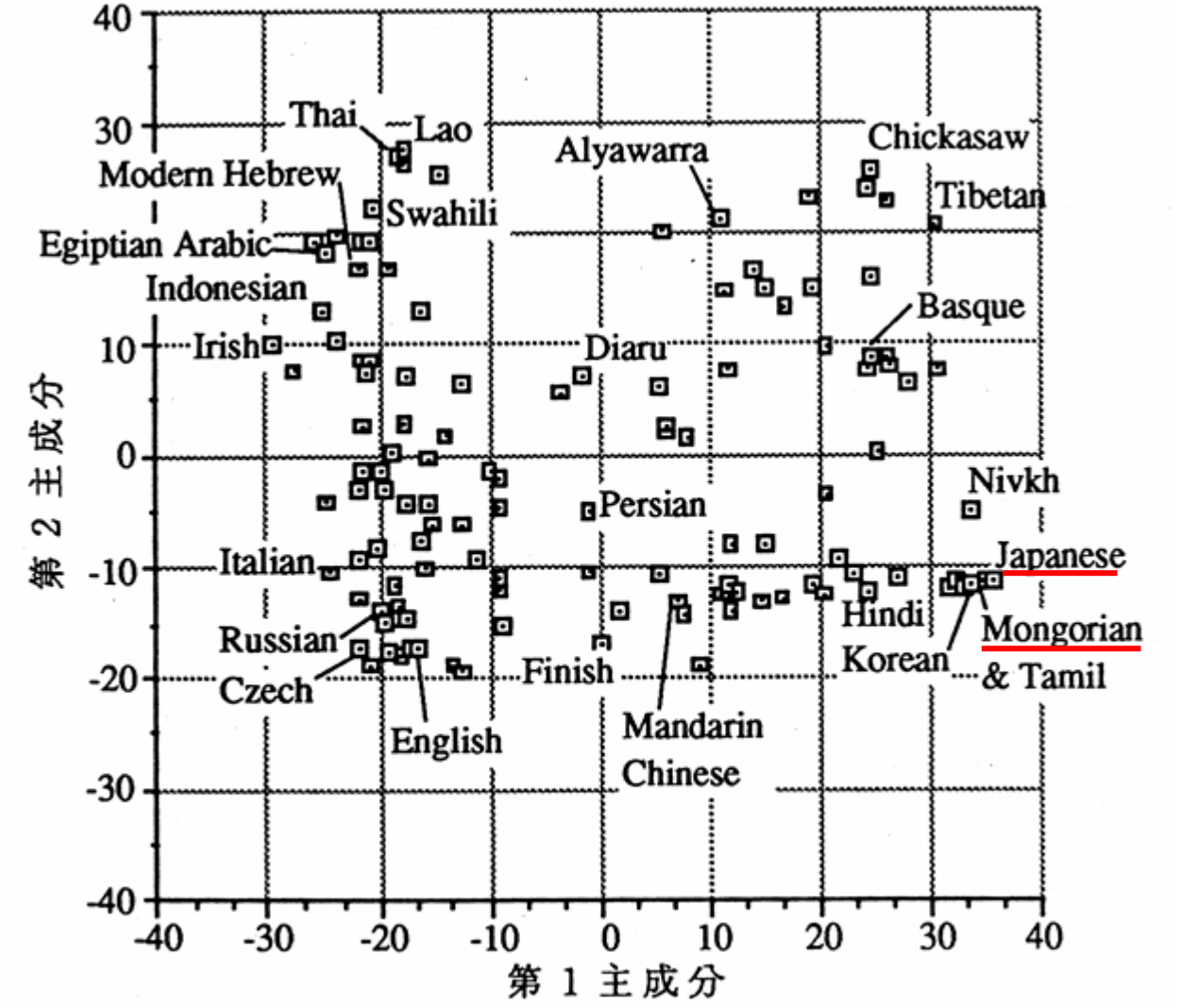
KIMURA Nobuyuki

* Tokyo University of Science, Suwa

eharate@rs.suwa.tus.ac.jp

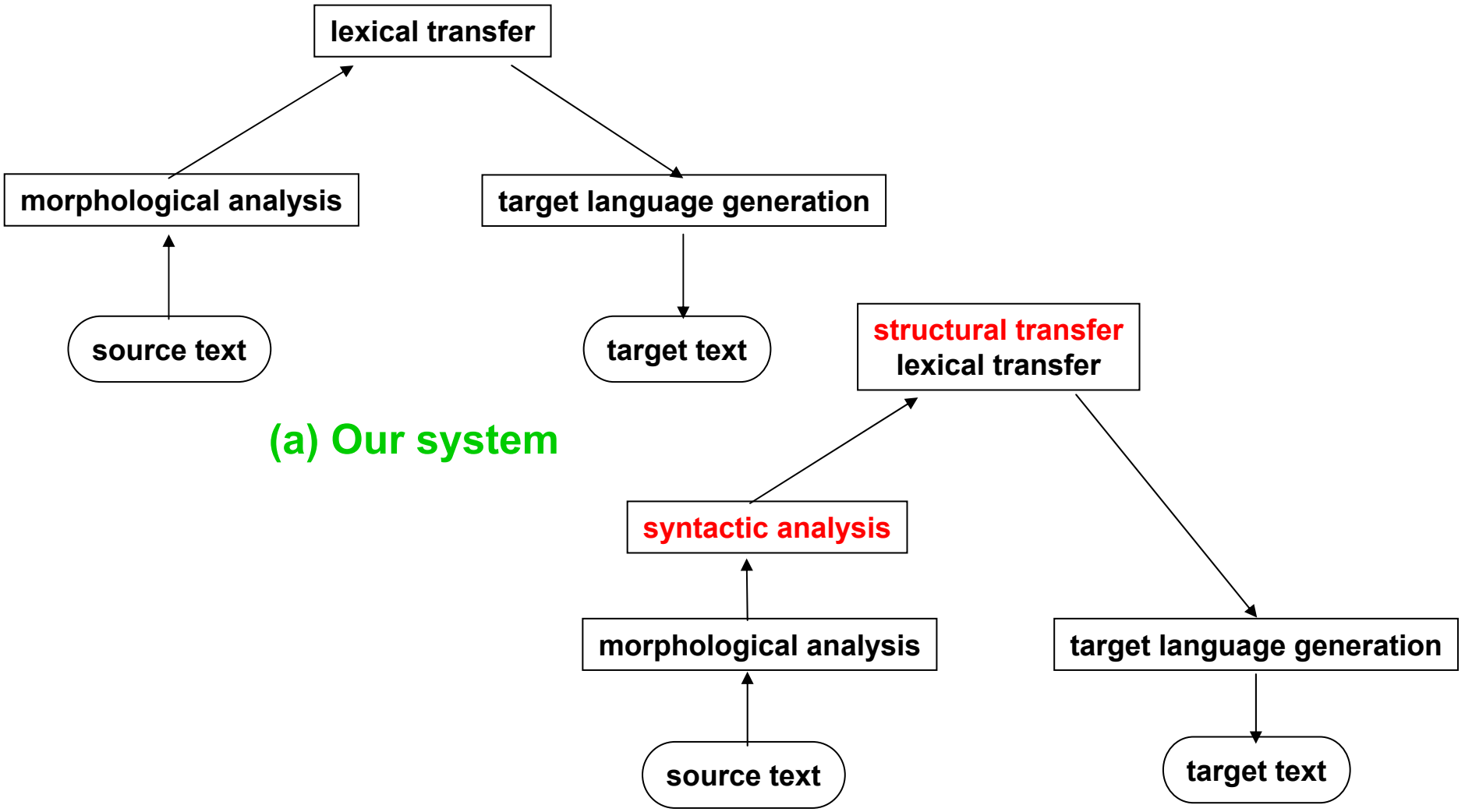
http://www.rs.suwa.tus.ac.jp/eharate

Distribution of languages in the two dimensional word order space



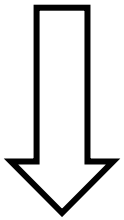
Ehara, Terumasa: Relation among word order parameters analyzed by multi-dimensional scaling, Proc. of the first annual meeting of the Association for Natural Language Processing, pp.173—176, Mar., 1995 (in Japanese) http://www.rs.suwa.tus.ac.jp/eharate/ehara/ronbun.files/relation_among_word_order_parameters_ENG.pdf.

System architecture



Mongolian new character set expressed by Japanese EUC

Japanese EUC Russian Cyrillic alphabet + Ө and Ү



Japanese EUC Russian Cyrillic alphabet
+
Greece Θ and English V

Language resources for ChaSen

- Grammar table (grammar.cha)
- Conjugate type table (ctypes.cha)
- Conjugate form table (cforms.cha)
- Connection matrix (connect.cha)

- Content word dictionary
- Function word dictionary

Grammar hierarchy (part of speech hierarchy)

Top level

verb
verb without stem conjugation
noun
noun without stem conjugation
adjective
adverb
conjunction
interjection
postposition
verb ending
noun ending
unknown word
word space
sentence finals
comma
opened parenthesis
closed parenthesis
number

Second level

masculine а
masculine о
feminine э
feminine *θ*
neuter

Third level

free form
affixes

4th level

vowel final
sonolant consonant final
hidden consonant final
unsonolant consonant final
ж ч ш final
ь final

5th level

vowel
sonolant м
sonolant н
sonolant г
sonolant л
sonolant б
sonolant в
sonolant р
hidden н
hidden г
unsonolant т
unsonolant з
unsonolant ц
unsonolant х
unsonolant ф
unsonolant д
unsonolant с
unsonolant к
unsonolant п
unsonolant ж
unsonolant ш
unsonolant ч
consonant ь

Conjugation type and conjugation form table

conjugation type	base form	c-form1	c-form2
hidden θ н	*	н	θ н
hidden а н	*	н	а н
hidden г	*	г	
hidden и н	*	н	и н
hidden н	*	н	н
hidden о н	*	н	о н
hidden э н	*	н	э н
sonolant v б	v б	б	
sonolant v в	v в	в	
sonolant v г	v г	г	
sonolant v л	v л	л	v
sonolant v м	v м	м	
sonolant v н	v н	н	v
sonolant v р	v р	р	v
sonolant θ б	θ б	б	
sonolant θ в	θ в	в	
sonolant θ г	θ г	г	
sonolant θ л	θ л	л	θ
sonolant θ м	θ м	м	
sonolant θ н	θ н	н	θ
sonolant θ р	θ р	р	θ
sonolant а б	а б	б	
sonolant а в	а в	в	
sonolant а г	а г	г	
sonolant а л	а л	л	а
sonolant а м	а м	м	
sonolant а н	а н	н	а
sonolant а р	а р	р	а
sonolant и б	и б	б	
sonolant и в	и в	в	
sonolant и г	и г	г	
sonolant и л	и л	л	и
sonolant и м	и м	м	
sonolant и н	и н	н	и
sonolant и р	и р	р	и

conjugation type	base form	c-form1	c-form2
sonolant л	л	л	*
sonolant н	н	н	*
sonolant о б	о б	б	
sonolant о в	о в	в	
sonolant о г	о г	г	
sonolant о л	о л	л	о
sonolant о м	о м	м	
sonolant о н	о н	н	о
sonolant о р	о р	р	о
sonolant р	р	р	*
sonolant у б	у б	б	
sonolant у в	у в	в	
sonolant у г	у г	г	
sonolant у л	у л	л	у
sonolant у м	у м	м	
sonolant у н	у н	н	у
sonolant у р	у р	р	у
sonolant э б	э б	б	
sonolant э в	э в	в	
sonolant э г	э г	г	
sonolant э л	э л	л	э
sonolant э м	э м	м	
sonolant э н	э н	н	э
sonolant э р	э р	р	э
vowel v й	v й	v	*
vowel θ й	θ й	θ	*
vowel а й	а й	а	*
vowel е й	е й	е	*
vowel ё й	ё й	ё	*
vowel о й	о й	о	*
vowel у й	у й	у	*
vowel э й	э й	э	*
vowel я й	я й	я	*
consonant ь	ь	*	и
consonant я	я	я	

conjugation type	base form	c-form1	c-form2
vowel v	v	*	
vowel θ	θ	*	
vowel а	а	*	
vowel и	и	*	
vowel о	о	*	
vowel у	у	*	
vowel э	э	*	
vowel и й	и й	и	*
unsonolant v д	v д	д	*
unsonolant v ж	v ж	ж	
unsonolant v з	v з	з	
unsonolant v с	v с	с	
unsonolant v т	v т	т	
unsonolant v ц	v ц	ц	
unsonolant v ч	v ч	ч	
unsonolant v ш	v ш	ш	
unsonolant θ д	θ д	д	*
unsonolant θ ж	θ ж	ж	
unsonolant θ з	θ з	з	
unsonolant θ с	θ с	с	
unsonolant θ т	θ т	т	
unsonolant θ ц	θ ц	ц	
unsonolant θ ч	θ ч	ч	
unsonolant θ ш	θ ш	ш	
unsonolant а д	а д	д	*
unsonolant а ж	а ж	ж	
unsonolant а з	а з	з	
unsonolant а с	а с	с	
unsonolant а т	а т	т	
unsonolant а ц	а ц	ц	
unsonolant а ч	а ч	ч	
unsonolant а ш	а ш	ш	
unsonolant д	д	д	*
unsonolant и д	и д	д	*
unsonolant и ж	и ж	ж	

conjugation type	base form	c-form1	c-form2
unsonolant и з	и з	з	
unsonolant и с	и с	с	
unsonolant и т	и т	т	
unsonolant и х	и х	с	
unsonolant и ц	и ц	ц	
unsonolant и ч	и ч	ч	
unsonolant и ш	и ш	ш	
unsonolant о д	о д	д	*
unsonolant о ж	о ж	ж	
unsonolant о з	о з	з	
unsonolant о с	о с	с	
unsonolant о т	о т	т	
unsonolant о ц	о ц	ц	
unsonolant о ч	о ч	ч	
unsonolant о ш	о ш	ш	
unsonolant у д	у д	д	*
unsonolant у ж	у ж	ж	
unsonolant у з	у з	з	
unsonolant у с	у с	с	
unsonolant у т	у т	т	
unsonolant у ц	у ц	ц	
unsonolant у ч	у ч	ч	
unsonolant у ш	у ш	ш	
unsonolant э д	э д	д	*
unsonolant э ж	э ж	ж	
unsonolant э з	э з	з	
unsonolant э с	э с	с	
unsonolant э т	э т	т	
unsonolant э ц	э ц	ц	
unsonolant э ч	э ч	ч	
unsonolant э ш	э ш	ш	

* : no character

Connection matrix table

Number of entries of the table : 6,941

((((("verb" "masculine a " "free form")))
    (((“verb” “masculine a ” “verb to verb affix”)))) 300)
((((("verb" "masculine a " "free form")))
    (((“verb without stem conjugation” “masculine a ” “verb to verb affix”)))) 300)
((((("verb" "masculine a " "free form")))
    (((“noun” “masculine a ” “verb to noun affix”)))) 300)

A sample of connection table data

Dictionaries

- base form (entry)
- grammar categories (i.e. part-of-speech, from top to 5th level)
- conjugation type (only for verb and noun)
- Japanese translation (semantic information field)
- morpheme cost for the morphological analysis

dictionary items

Content word dictionary* : 9,467 entries
Function word dictionary : 1,635 entries

* Shimizu, Mikio: Electronic Japanese Mongolian word index (dictionary), 2001. http://mk-smz.hp.infoseek.co.jp//OnEJMD.shtml

Sample of content word dictionary entry

(“pos” (“noun” “feminine э ” “free form”
“sonolant consonant final”))
((“entry” (v з э г 1000))
(“conjugation type” “sonolant consonant э г ”)
(“semantic information” {ペン/ボールペン}))

(“pos” (“verb without stem conjugation” “feminine э ”
“free form” “unsonolant consonant final” “unsonolant с ”))
((“entry” (vv с 1000))
(“semantic information” {生まれる/生じる/成り立つ/発生する}))

Sample of function word dictionary entry

(“pos” (“noun ending” “feminine θ ”))
((“entry” (и й г 100))
(“semantic information” {〈普通格変化対格〉/【を】}))

(“pos” (“verb ending” “masculine o ”))
((“entry” (o x 100))
(“semantic information” {〈形動詞形__連体現在と未来〉/【】}))

Structural transfer

- minor word order change

  number of rules : 7

  example
  verb+past_tense+negation
  ⇒ verb+negation+past_tense
  м э д +с э н +г v й
  ⇒ 知る(м э д)+ない(г v й)+た(с э н)

Lexical transfer

- Compound word translation
  8 rules
  example
  verb+ж+э х л э х
  ⇒ verb+て(ж)+はじめる(э х л э х)
- Zero elements complement
  14 rules
  example
  noun+zero_case_marker
  ⇒ noun+{〈格助詞〉/【】/【の】/【と】/【が】/【は】/【を】/【に】/【で】}

Japanese generation

- translation selection
  statistical method
  simple translation model
  dependency type language model
  (10 years Japanese newspaper)

- conjugation form selection
  expansion and selection

Experimental results

<Input>

а х ы н х а а __ и р с н и й г __ м э д с э н г v й

<Output of the morphological analysis phase>

morpheme form|base form|POS|c-type|c-form|Japanese translation|cost|total cost

а х|а х|名詞不変 男性 a 自立 非共鳴子音語幹 х 語幹|||兄|1000|1500
ы н|ы н|名詞語尾 男性 a |||{〈普通格変化__属格〉/【の】}|100|1700
х а а|х а а|名詞語尾 男性 a |||{〈非人称再帰所有格変化__属格〉/自分【の】}|100|1900
__|__|空白|||__|1000|3000
и р|и р|動詞 女性 э 自立 共鳴子音語幹|共鳴子音 р|基本形|{来る/見える}|1000|4500
с н|с|名詞 女性 э 動詞ー名詞接辞 隠れ子音語幹|隠れ э н|活用形1|
{〈結果 с а н〉/【たもの】/【だもの】/【たこと】/【だこと】}|500|5100
и й г|и й г|名詞語尾 女性 э |||{〈普通格変化__対格〉/【を】}|100|5300
__|__|空白|||__|1000|6400
м э д|м э д|動詞 女性 э 自立 非共鳴子音語幹|非共鳴子音 д|基本形|
{感じる/知る【ている】/知る/解る}|1000|7900
с э н|с э н|動詞語尾 女性 э |||{〈形動詞形__連体過去〉/【た】/【だ】}|100|8100
г v й|г v й|動詞語尾 女性 э |||{〈形動詞形__連体否定〉/【ない】}|100|8300

<Output of the transfer phase>

|S|
|P|
|M| x a a | x a a |名詞語尾 男性 a |||{〈非人称再帰所有格変化__属格〉/自分【の】}|100|1900|/M|
|/P|
|P|
|M| a x | a x |名詞不変 男性 a 自立 非共鳴子音語幹 x 語幹|||兄|1000|1500|/M|
|M| ы н | ы н |名詞語尾 男性 a |||{〈普通格変化__属格〉/【の】}|100|1700|/M|
|/P|
|P|
|M| и р | и р |動詞 女性 э 自立 共鳴子音語幹|共鳴子音 р |基本形|{来る/見える}|1000|4500|/M|
|M| с н | с |名詞 女性 э 動詞ー名詞接辞 隠れ子音語幹|隠れ э н |活用形1|
{〈結果 с а н〉/【たもの】/【だもの】/【たこと】/【だこと】}|500|5100|/M|
|M| и й г | и й г |名詞語尾 女性 э |||{〈普通格変化__対格〉/【を】}|100|5300|/M|
|/P|
|P|
|M| м э д | м э д |動詞 女性 э 自立 非共鳴子音語幹|非共鳴子音 д |基本形|
{感じる/知る【ている】/知る/解る}|1000|7900|/M|
|M| г v й | г v й |動詞語尾 女性 э |||{〈形動詞形__連体否定〉/【ない】}|100|8300|/M|
|M| с э н | с э н |動詞語尾 女性 э |||{〈形動詞形__連体過去〉/【た】/【だ】}|100|8100|/M|
|/P|
|/S|

<Output of the translation selection phase>

|S|
|P| 0| 3|
|M| x a a | x a a |名詞語尾 男性 a |||自分【の】|〈非人称再帰所有格変化__属格〉|/M|
|/P|
|P| 1| 3|
|M| a x | a x |名詞不変 男性 a 自立 非共鳴子音語幹 x 語幹|||兄||/M|
|M| ы н | ы н |名詞語尾 男性 a |||【の】|〈普通格変化__属格〉|/M|
|/P|
|P| 2| 3|
M	и р	и р	動詞 女性 э 自立 共鳴子音語幹	共鳴子音 р	基本形	来る		/M
M	с н	с	名詞 女性 э 動詞ー名詞接辞 隠れ子音語幹	隠れ э н	活用形1	【たもの】	〈結果 с а н〉	/M
M	и й г	и й г	名詞語尾 女性 э			【を】	〈普通格変化__対格〉	/M
/P
P	3	-1	0
M	м э д	м э д	動詞 女性 э 自立 非共鳴子音語幹	非共鳴子音 д	基本形	知る【ている】		/M
M	г v й	г v й	動詞語尾 女性 э			【ない】	〈形動詞形__連体否定〉	/M
M	с э н	с э н	動詞語尾 女性 э			【た】	〈形動詞形__連体過去〉	/M
/P
/S

<Output of the generation phase>

自分の兄の来たものを知っていなかった

<Correct translation>

自分の兄が来たことを知らなかった

Conclusion

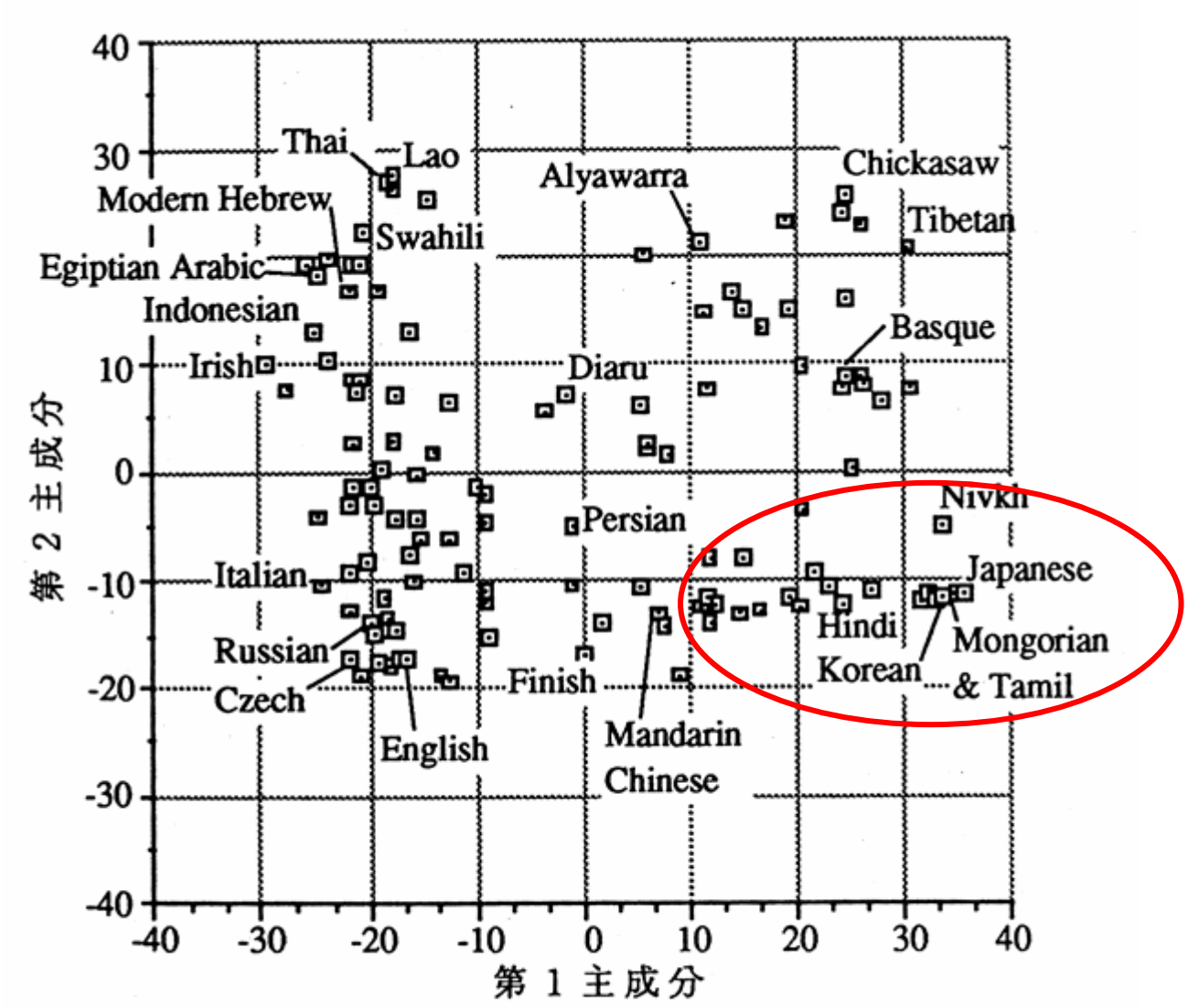
Present the structure of the translation system
from Mongolian to Japanese
which is based on Japanese morphological analyzer ChaSen.

Some translation results are also presented.

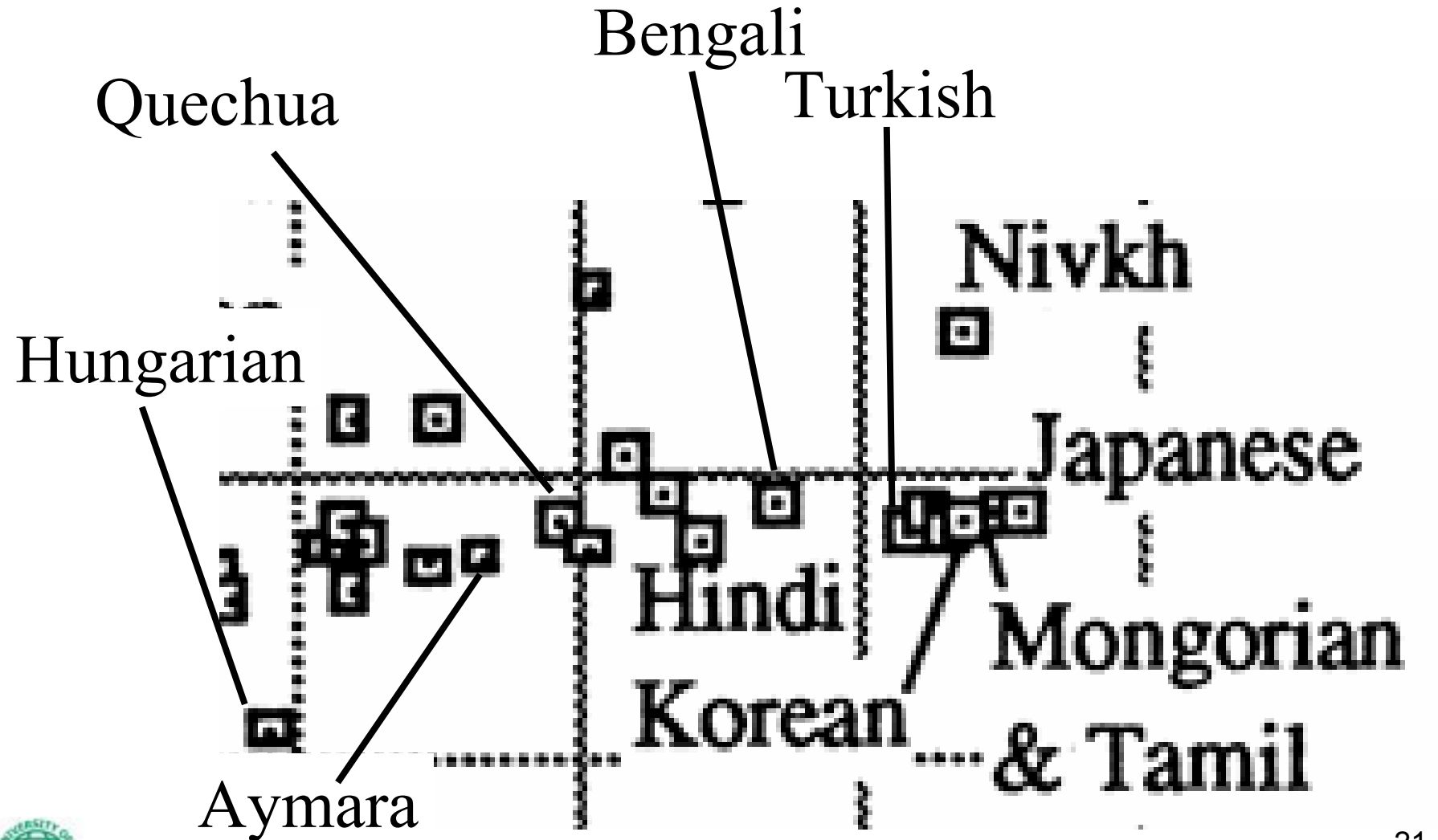
Future plan

- Construction of a large scale dictionary
- Using MJ bilingual corpus to brush up the system
- Extension the method to other languages

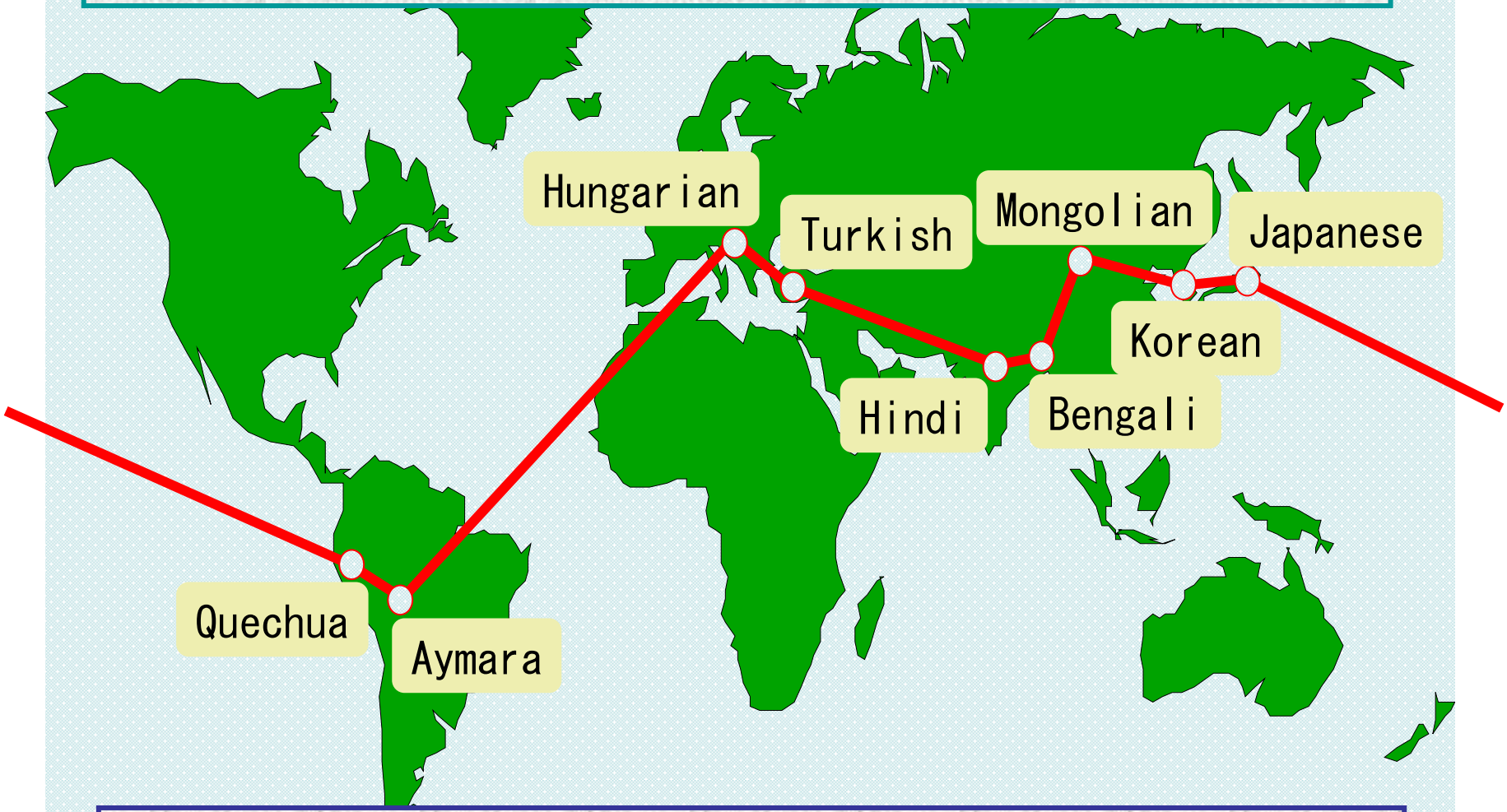
Distribution of languages in the two dimensional word order space



Languages that have similar word order with Japanese



Thank you for your attention



http://www.rs.suwa.tus.ac.jp/eharate