Part-of-speech tagging for a Southern Min Corpus

Ching Chu Sun (June Sun)

Department of Linguistics
University of Alberta

American Association for Corpus Linguistics AACL 2009

Southern Min
A Chinese language

Chinese Corpora (McEnery & Xiao)

Corpus	POS	Channel	Variety
LCMC	Yes	Written	China
Sinica	Yes	Mixed	Taiwan
PH	No	Written	China
PFR	Yes	Written	China
LIVAV	No	Written	Mixed
SCCSD	No	Spoken	China
TREC	No	Written	China
Gigaword	No	Written	China
Callhome	No	Spoken	Mixed

No (tagged) corpus of Southern Min is publicly available.

Data sources

The data used in the current study:

Written: http://taigu.fhl.net/index.html

Spoken: Part of the face-to face conversations between 3 native speakers of Southern Min

HunPos

(Halácsy, Kornai, & Oravecz 2007)

- HunPos: Finding the most probable sequence of tags for each sentence

- Formula: Calculating the probability of a given sequence of tags

$$\underset{t_1 \ldots t_T}{\operatorname{argmax}} P(t_{T+1} \mid t_T) \prod_{i=1}^{T} P(t_i \mid t_{i-1}, t_{i-2})\, P(w_i \mid t_{i-1}, t_i)$$

HunPos
(Halácsy, Kornai, & Oravecz 2007)

The procedure for finding probability of a tag sequence for a sentence:

(1) Product of (A) and (B)

(A) p(this tag|prev. 2 tags)

(B) p(this word|this tag & prev.tag)

(2) Multiply p(end tag|last tag of sequence)

By using HunPos for tagging POS, I'd like to find out…

- Will there be any significant difference in accuracy for tagging POS in romanized (phonetic) tokens and in Chinese character tokens?
- Will the accuracy of tagging increase with more training data?
- How accurate is the tagger with unseen words (new type words)?
- What tags is the tagger most and least accurate with?

Written

Romanization		Characters	
0	NUM	0	NUM
hai2to2	N	海島	N
tai5oan5	PW	台灣	PW
e5	AM	e5	AM
tek8sek4	N	特色	N
tai5oan5	PW	台灣	PW
si7	COP	是	COP
chit4	NUM	一	NUM
e5	CF	e5	CF
hai2to2	N	海島	N
,	PUNC	,	PUNC
i1	PN	伊	PN
e5	PM	e5	PM
tek8cheng4	N	特徵	N
toh5	ADV	就	ADV
si7	COP	是	COP
hai2hoaN7	N	海岸線	N
chin4	INT	真	INT
tng5	SV	長	SV
,	PUNC	,	PUNC

Spoken

Romanization		Characters	
0	NUM	0	NUM
tiam3	PREP	佇	PREP
tai5oan5	PW	台灣	PW
long2	ADV	long2	ADV
si7	COP	是	COP
kong2	V	講	V
kok4gi3	N	國語	N
ma7	PMOD	0	NUM
khah4	INT	ma7	PMOD
be7	AV	khah4	INT
		欲	AV
a2	FR	a2	FR
m7ku2	CCL	m7ku2	CCL
chhiuN4	V	像	V
goa2	PNPERS	我	PNPERS
hit8	DEM	彼	DEM
kang4	TW	kang4	TW
lok8	V	錄	V
nng7	NUM	兩	NUM
e5	CF	e5	CF
u5bo5	V	有無	V
hoN1	QM	hoN1	QM

Methodology

	Written (1462)		Spoken (1425)	
	Romanization	Character	Romanization	Character
set 1	278	278	244	244
set 2	141	141	117	117
set 3	146	146	134	134
set 4	149	149	123	123
set 5	94	94	126	126
set 6	110	110	120	120
set 7	106	106	115	115
set 8	106	106	134	134
set 9	156	156	134	134
set 10	176	176	178	178

total tokens

tagged 1184 (wr, wh) 1181 (sr, sh)

Methodology

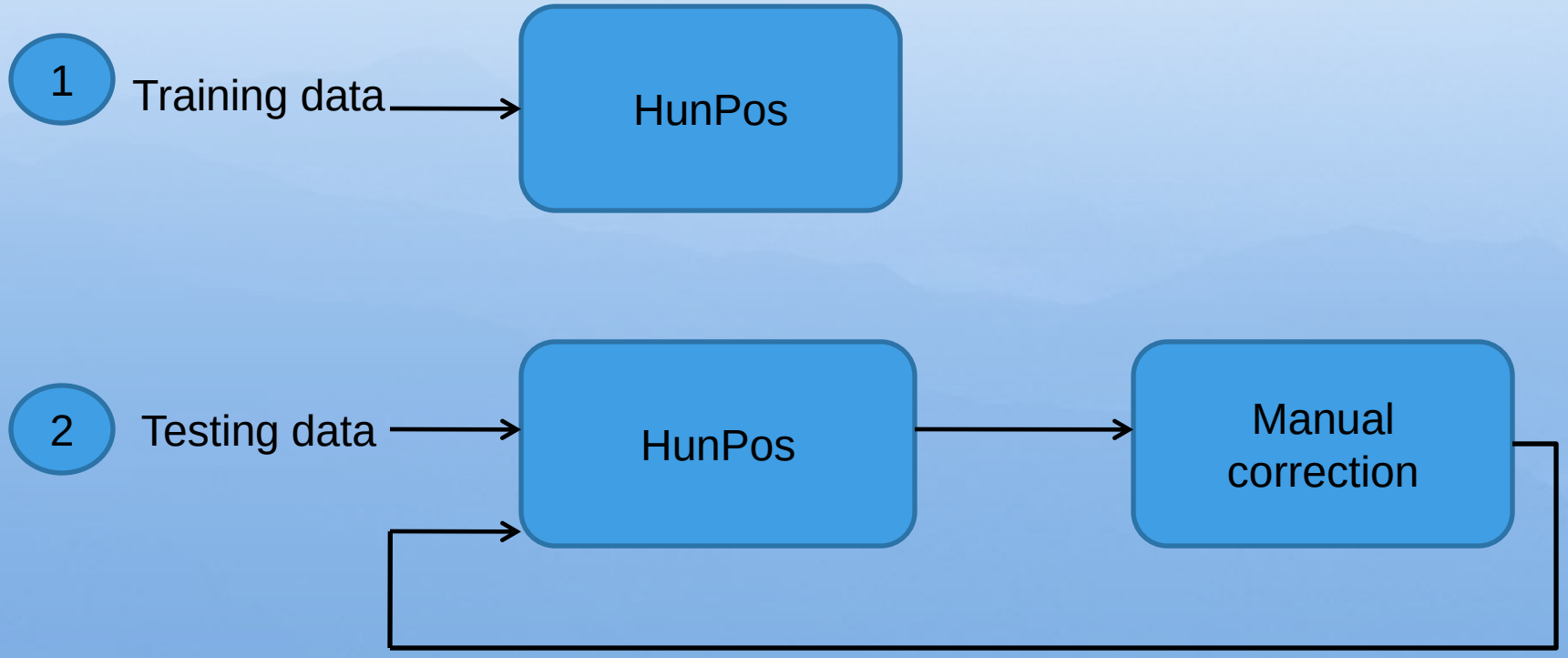
- Manually POS tagging principle

(1) Using “A Dictionary of Southern Min” by Embree (1984) as a starting point

(2) Customized tags

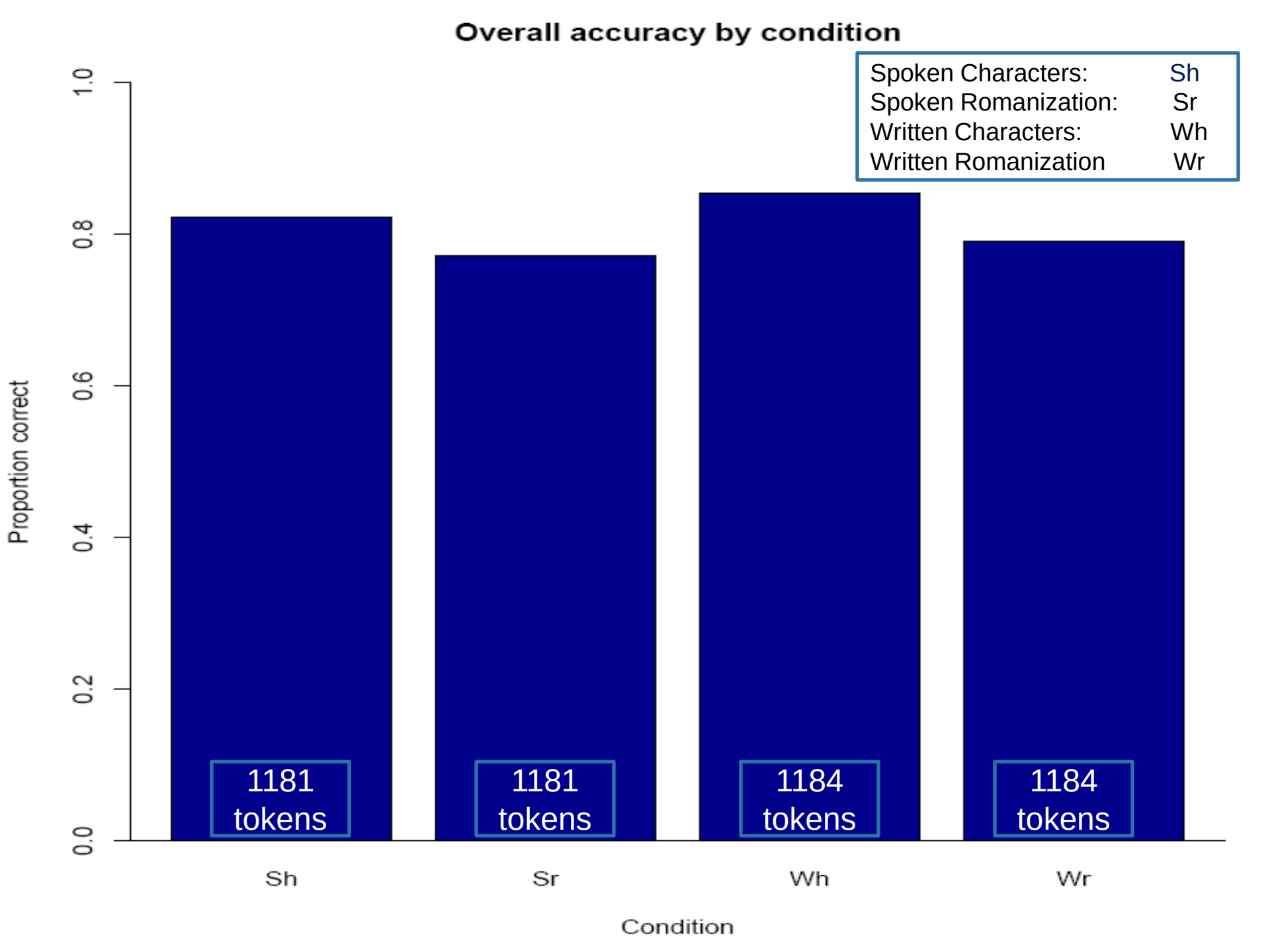
- Without default tags
- A new tag was added when necessary

Procedure of training the tagger



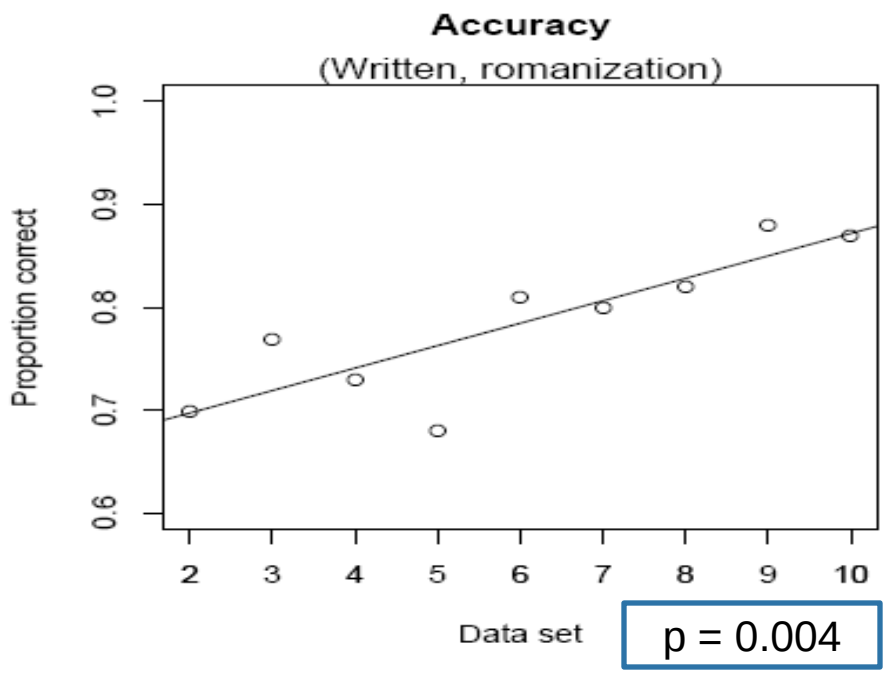
Question 1:

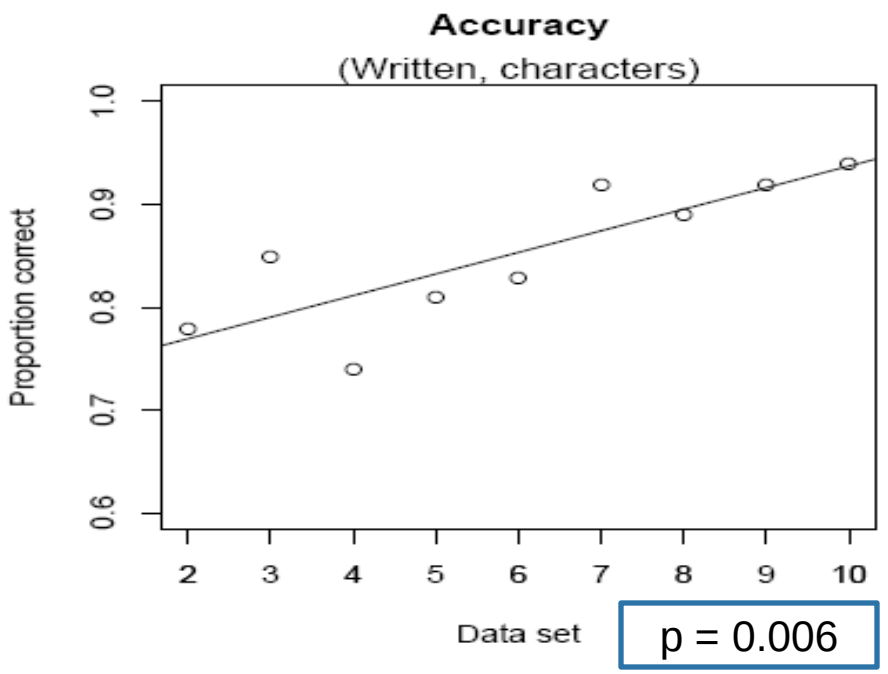
Will there be any significant difference in accuracy for tagging POS in romanized tokens and in Chinese character tokens?

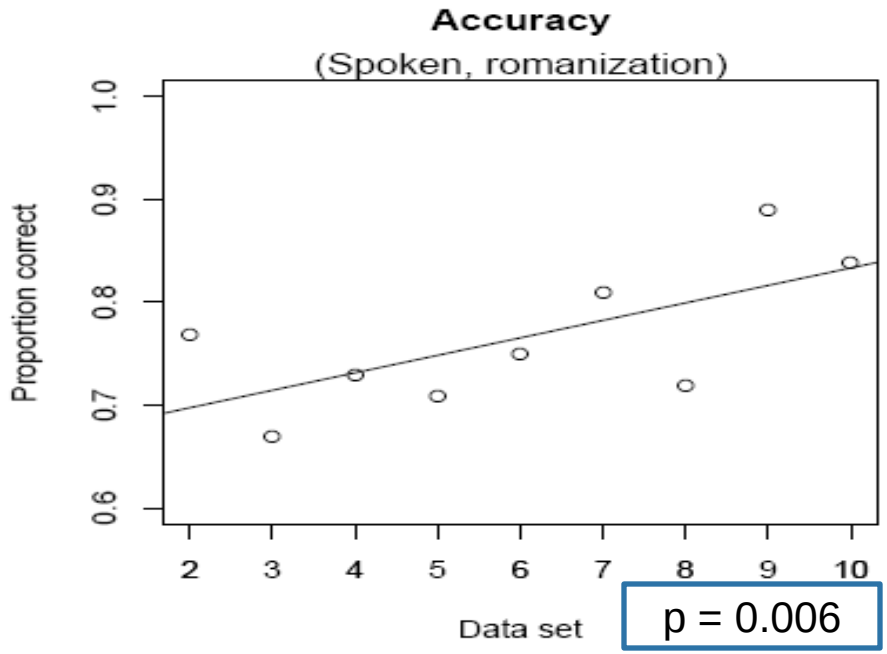

Overall accuracy by condition
Spoken Characters: Sh
Spoken Romanization: Sr
Written Characters: Wh
Written Romanization Wr
Proportion correct
0.0
0.2
0.4
0.6
0.8
1.0
1181 tokens
1181 tokens
1184 tokens
1184 tokens
Sh
Sr
Wh
Wr
Condition

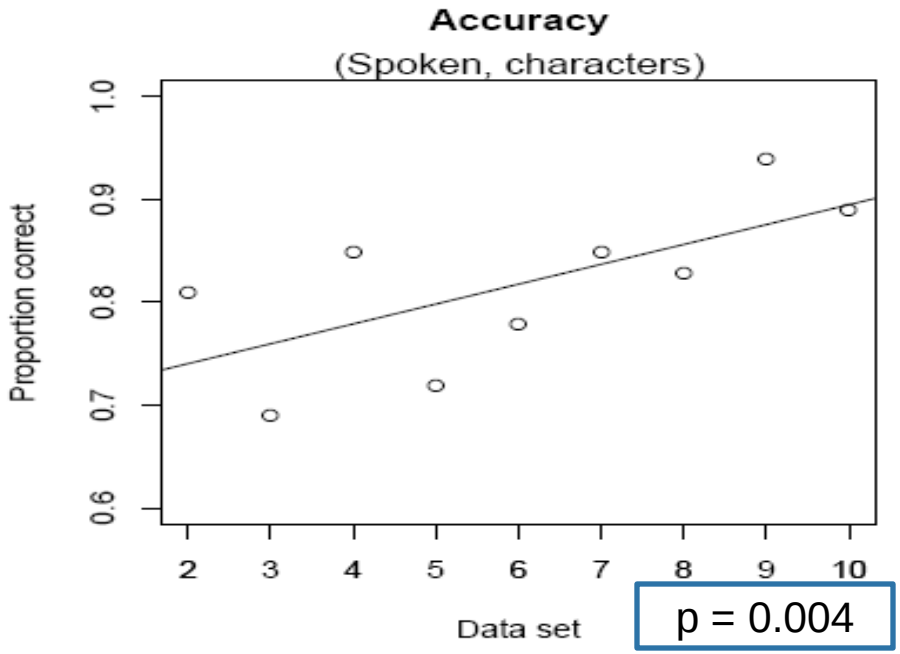
Question 2:

Will the accuracy of tagging increase with more training data?


Accuracy
(Written, romanization)
Proportion correct
1.0
0.9
0.8
0.7
0.6
2
3
4
5
6
7
8
9
10
Data set
p = 0.004

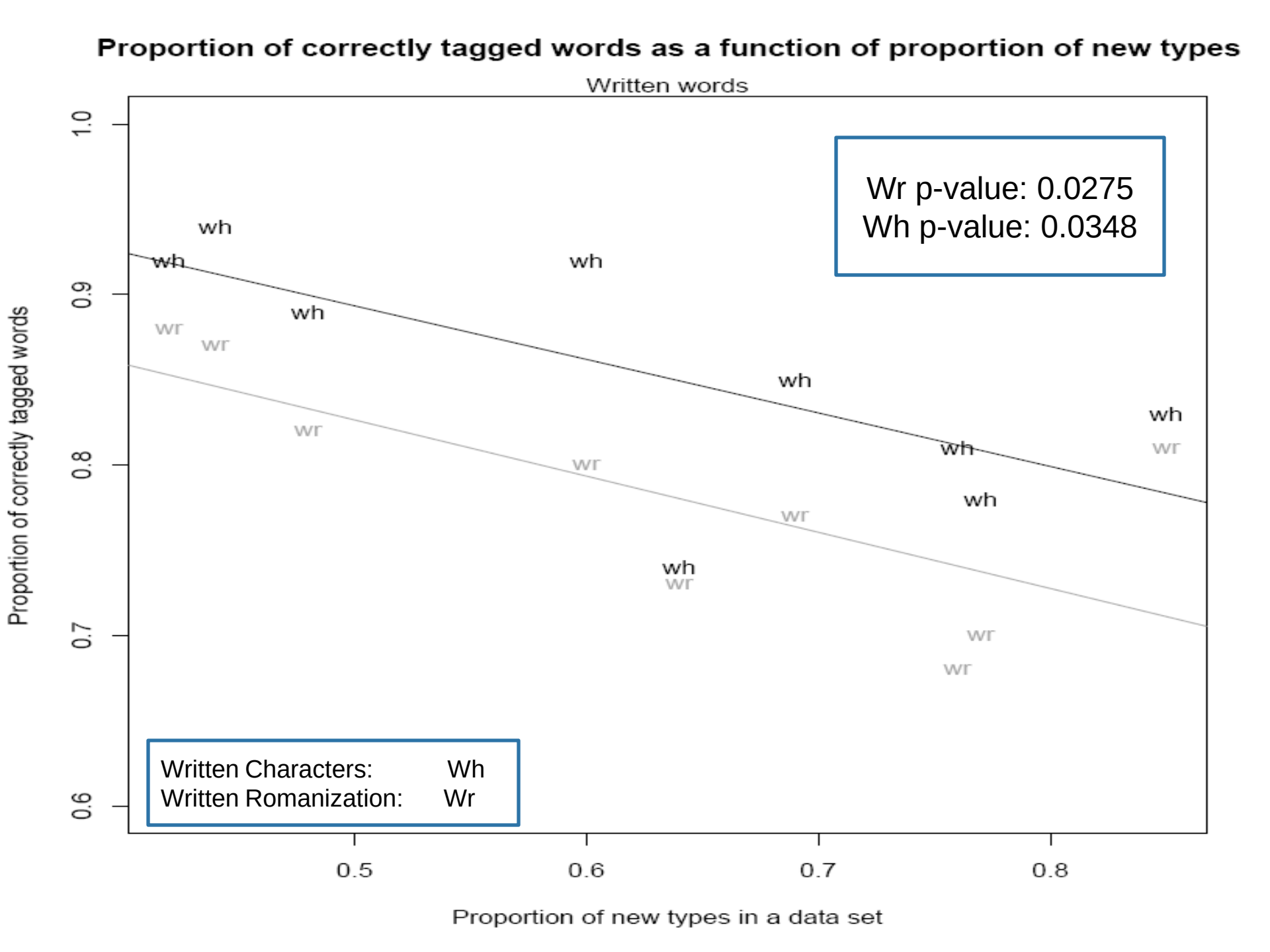

Accuracy
(Written, characters)
Proportion correct
1.0
0.9
0.8
0.7
0.6
2
3
4
5
6
7
8
9
10
Data set
p = 0.006

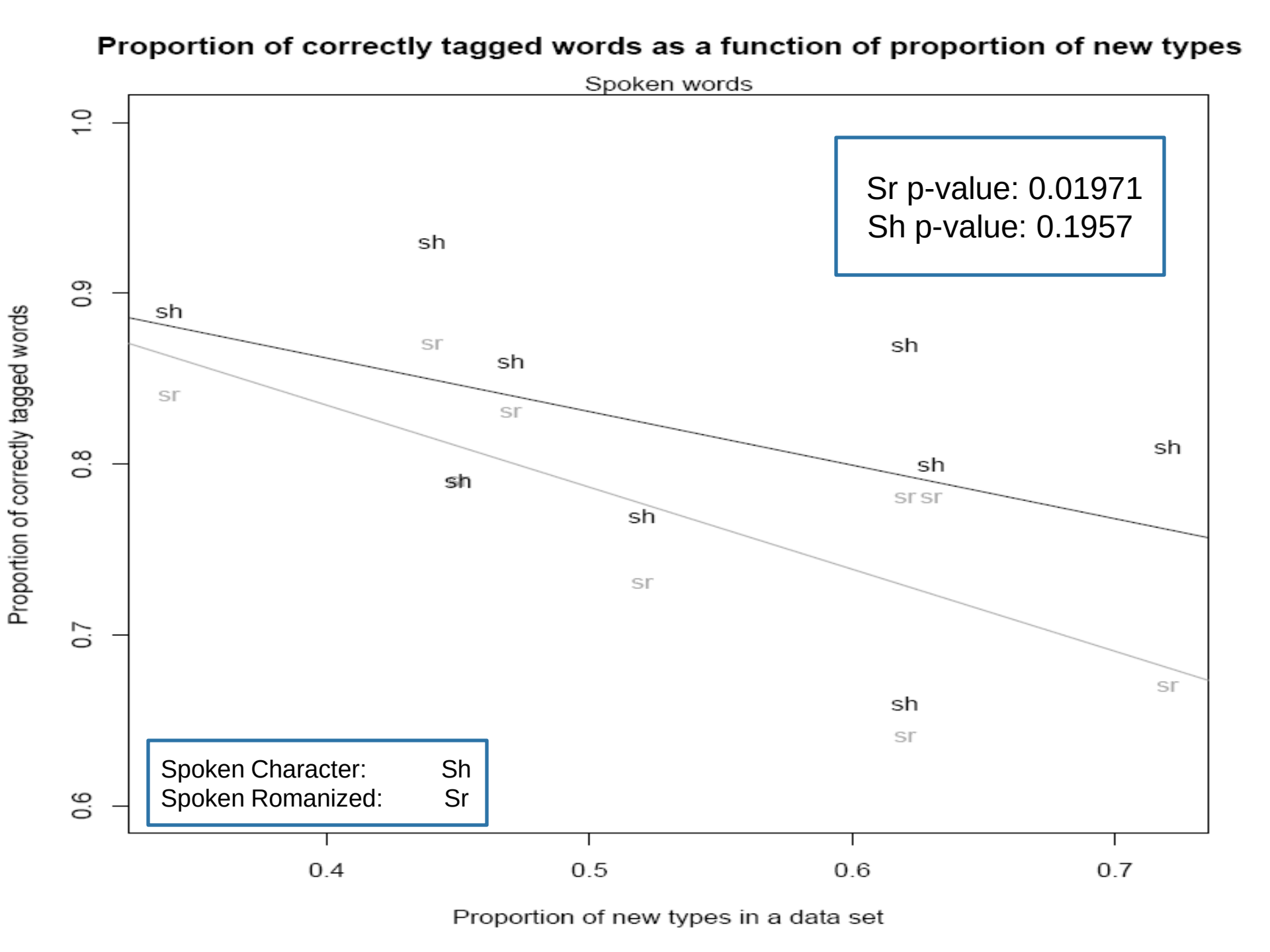

Accuracy
(Spoken, romanization)
Proportion correct
1.0
0.9
0.8
0.7
0.6
2
3
4
5
6
7
8
9
10
Data set
p = 0.006


Accuracy
(Spoken, characters)
Proportion correct
1.0
0.9
0.8
0.7
0.6
2
3
4
5
6
7
8
9
10
Data set
p = 0.004

Question 3:

How accurate is the tagger with unseen words?


Proportion of correctly tagged words as a function of proportion of new types
Written words
Wr p-value: 0.0275
Wh p-value: 0.0348
Written Characters: Wh
Written Romanization: Wr
Proportion of correctly tagged words
Proportion of new types in a data set
1.0
0.9
0.8
0.7
0.6
0.5
0.6
0.7
0.8
wh
wr

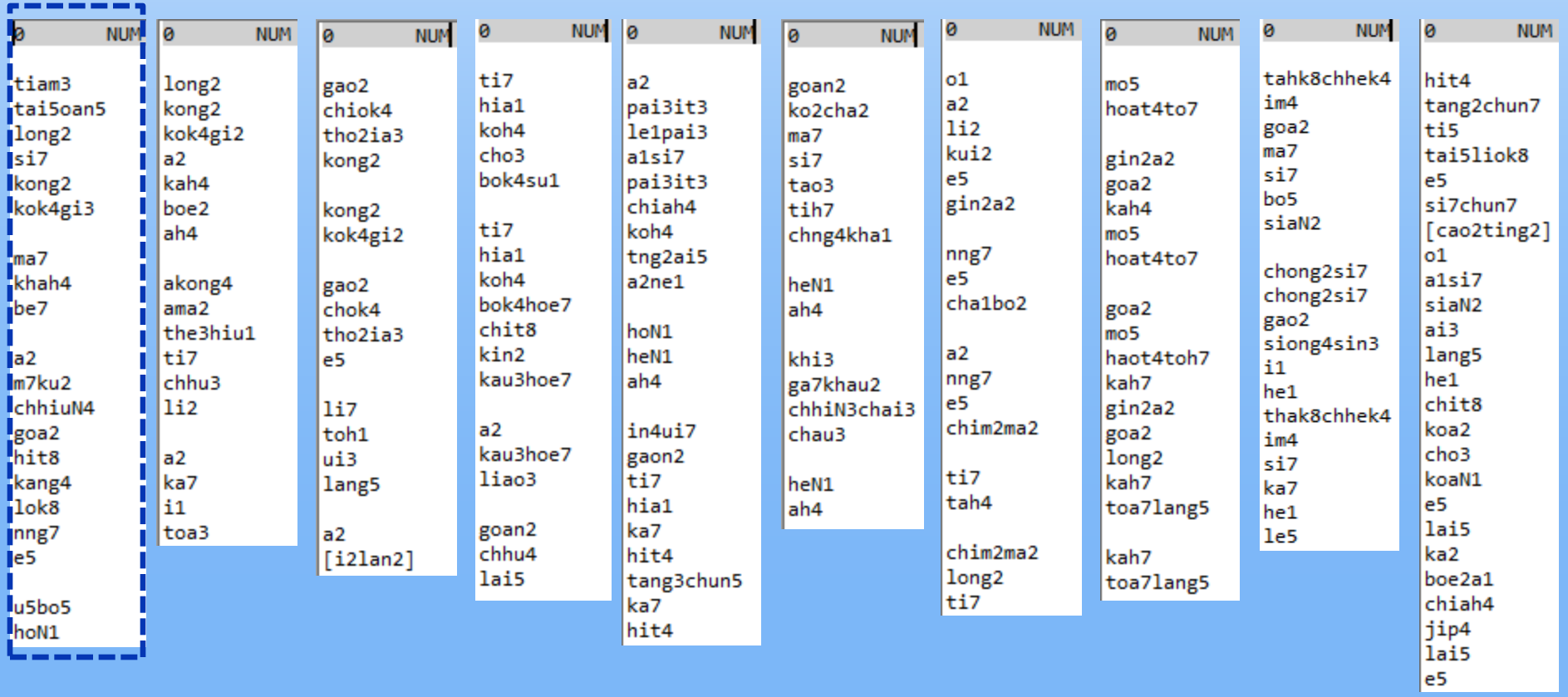

Proportion of correctly tagged words as a function of proportion of new types
Spoken words
Sr p-value: 0.01971
Sh p-value: 0.1957
Spoken Character: Sh
Spoken Romanized: Sr
Proportion of correctly tagged words
Proportion of new types in a data set
1.0
0.9
0.8
0.7
0.6
0.4
0.5
0.6
0.7

Question 4:

What tags is the tagger most and least accurate with?

Evaluating tagging accuracy of individual tags by cross validation

Leave-1-Out cross validation



0	NUM
tiam3	
tai5oan5	
long2	
si7	
kong2	
kok4gi3	
ma7	
khah4	
be7	
a2	
m7ku2	
chhiuN4	
goa2	
hit8	
kang4	
lok8	
nng7	
e5	
u5bo5	
hoN1	

0	NUM
long2	
kong2	
kok4gi2	
a2	
kah4	
boe2	
ah4	
akong4	
ama2	
the3hiu1	
ti7	
chhu3	
li2	
a2	
ka7	
i1	
toa3	

0	NUM
gao2	
chiok4	
tho2ia3	
kong2	
kong2	
kok4gi2	
gao2	
chok4	
tho2ia3	
e5	
li7	
toh1	
ui3	
lang5	
a2	
[i2lan2]	

0	NUM
ti7	
hia1	
koh4	
cho3	
bok4su1	
ti7	
hia1	
koh4	
bok4hoe7	
chit8	
kin2	
kau3hoe7	
a2	
kau3hoe7	
liao3	
goan2	
chhu4	
lai5	

0	NUM
a2	
pai3it3	
le1pai3	
a1si7	
pai3it3	
chiah4	
koh4	
tng2ai5	
a2ne1	
hoN1	
heN1	
ah4	
in4ui7	
gaon2	
ti7	
hia1	
ka7	
hit4	
tang3chun5	
ka7	
hit4	

0	NUM
goan2	
ko2cha2	
ma7	
si7	
tao3	
tih7	
chng4kha1	
heN1	
ah4	
khi3	
ga7khau2	
chhiN3chai3	
chau3	
heN1	
ah4	

0	NUM
o1	
a2	
li2	
kui2	
e5	
gin2a2	
nng7	
e5	
cha1bo2	
a2	
nng7	
e5	
chim2ma2	
ti7	
tah4	
chim2ma2	
long2	
ti7	

0	NUM
mo5	
hoat4to7	
gin2a2	
goa2	
kah4	
mo5	
hoat4to7	
goa2	
mo5	
haot4toh7	
kah7	
gin2a2	
goa2	
long2	
kah7	
toa7lang5	
kah7	
toa7lang5	

0	NUM
tahk8chhek4	
im4	
goa2	
ma7	
si7	
bo5	
siaN2	
chong2si7	
chong2si7	
gao2	
siong4sin3	
i1	
he1	
thak8chhek4	
im4	
si7	
ka7	
he1	
le5	

0	NUM
hit4	
tang2chun7	
ti5	
tai5liok8	
e5	
si7chun7	
[cao2ting2]	
o1	
a1si7	
siaN2	
ai3	
lang5	
he1	
chit8	
koa2	
cho3	
koaN1	
e5	
lai5	
ka2	
boe2a1	
chiah4	
jip4	
lai5	
e5	

Result from cross validation for individual tags

Spoken romanization

Total tags:31			Correct prop.
copula (41) progressive marker (5)	final particle(27) post clause modifier (3)	conjunction (8)	100%
personal pronoun (100/104) specifier (20/22) negation1 (18/19)	clause connector (18/20) **attributive marker (18/19)** negation 2(11/12)		90%-99%
verb (159/185) numeral (24/29) intensifier (14/17) discourse marker (8/10)	adverb (64/76) preposition (20/25) question marker(10/12) perfective marker (7/8)	filler (38/44)	80%-89%
noun (108/136) auxiliary verb (15/21)	predicate modifier (17/23)		70%-79%
place word(25/47)	classifier (19/33)	time word (10/19)	50-59%
stative verb (12/32) location direction(2/5)	pronoun (5/11) numeral modifier(0/1)		Below 50%

Result from cross validation for individual tags

Spoken characters

Total tags: 31			Correct prop.
copula (41) post clause modifier (3)	determiner (3) progressive marker (5)	final particle (27)	100%
attributive marker (18/19) personal pronoun (93/103)	negation 2(11/12)		90%-99%
verb (151/185) specifier (19/22) question marker (10/12) conjunction(7/8)	adverb (63/76) clause connector (17/20) discourse marker (8/10)	filler (39/44)	80%-89%
noun (105/136) preposition (18/24) intensifier (12/17)	numeral (22/29) predicate modifier (18/23)		70%-79%
place word (29/47) auxiliary verb (14/21) location direction (3/5)	classifier (20/33) perfective marker (5/8)		60%-69%
time word (11/19)			50-59%
stative verb (12/32) numeral modifier (0/1)	negation 1(9/19)	pronoun (4/12)	Below 50%

Result from cross validation for individual tags

Written romanization

Total tags: 29			Correct prop.
sentence punctuation (24) numeral modifier (2) predicate modifier (2)	**copula (16)** classifier (2)	negation1 (4)	100%
attributive marker (62/64) intensifier (9/10)	place word (31/32)		90%-99%
noun (161/199) conjunction (10/12)	punctuation (124/142) time word (5/6)		80%-89%
adverb (49/63) clause connector (10/13)	preposition (14/18) post verb predicate (6/8)		70%-79%
verb (90/133) pronoun (16/24)	numeral (20/31)		60%-69%
stative verb (24/46) specifier (1/2)	negation 2 (1/2)		50-59%
location direction (5/14) personal pronoun(2/5) noun suffix (0/1)	auxiliary verb (4/12) determiner (0/1) pronoun suffix (0/1)		Below 50%

Result from cross validation for individual tags

Written characters

Total tags: 29			Correct prop.
sentence punctuation (25) numeral modifier (2)	**copula (16)** predicate modifier (2)	negation 1(4)	100%
attributive marker (63/64) punctuation(130/142)	intensifier (9/10) place word (29/32)		90%-99%
noun (160/199)	time word (5/6)		80%-89%
conjunction (9/12)	preposition (14/18)		70%-79%
verb (84/133) pronoun (15/24)	**adverb (42/63)** clause connector (8/13)		60%-69%
auxiliary verb (6/12) location direction (8/14) specifier (1/2)	classifier (1/2) negation 2 (1/2)		50-59%
stative verb (18/46) post verb predicate (2/6) determiner (0/1) pronoun suffix (0/1)	numeral (15/31) personal pronoun (2/5) noun suffix (0/1)		Below 50%

Tagging difficulty

	Spoken character	Spoken romanization	Written character	Written romanization
80%-89%	Verb (151/185)	Verb (159/185)	Noun (161/199)	Noun (160/199)
70%-79%	Noun (105/136)	Noun (108/136)		
60%-69%			Verb (90/133)	Verb (84/133)
Below 50%	Stative verb (12/32)	Stative verb (12/32)	Stative verb (24/46)	Stative verb (18/46)

Tagging difficulty
Easy for manual labor but difficult for HunPos

Stative verb

	HunPos	me
goa2	PNPERS	PNPERS
siuN7	N	V
ka7	CONJ	CONJ
he7	N	SP
kang5khuang3	N	SV
lah4	DM	DM

	HunPos	me
goa2	PNPERS	PNPERS
kam2kak4	V	V
chiok4	INT	INT
su1si4	N	SV
e5	AM	AM
kong2	V	V

	HunPos	me
li2	PNPERS	PNPERS
na7	ADV	PMOD
ti7	PREP	PREP
hak1hau7	N	N
khah4	INT	INT
chiap8	V	SV
kong2	V	V
kok4gu2	N	N
hoN1	QM	QM

	HunPos	me
ma7	PMOD	PMOD
si7	COP	COP
be7	AV	AV
lian2tng2	V	SV

	HunPos	me
gao2	PNPERS	PNPERS
chiok4	ADV	INT
tho2ia3	V	SV
kong2	V	V

	HunPos	me
so2i2	CCL	CCL
li2	PNPERS	PNPERS
su7lim5	V	PW
chian7	TW	INT
sek4	V	SV

Tagging difficulty
Easy for manual labor but difficult for HunPos

Noun

	HunPos	me
akong4	TW	N
ama2	TW	N
the3hiu1	V	V
ti7	PREP	PREP
chhu3	N	N
li2	PNPERS	DL

	HunPos	me
[Winnie]	N	PN
si7	COP	COP
ti7	PREP	PREP
tai5oan5	PW	PW
to2	ADV	ADV
bo5	NEG1	NEG1
kong2	V	V
tai5gi3	V	N
ah4	FP	FP

	HunPos	me
chhiuN1	QM	V
gaon2	PNPERS	PNPERS
ti7ti7	CCL	N
in4	PNPERS	PNPERS
giaN2	V	N

Tagging difficulty
Easy for manual labor but difficult for HunPos

Verb

	HunPos	me
goa2	PNPERS	PNPERS
siuN7	N	V
ka7	CONJ	CONJ
he7	N	SP
kang5khuang3	N	SV
lah4	DM	DM

	HunPos	me
m7	V	NEG2
bat4	ADV	AV
u5	V	V
tui3	PREP	PREP
goa7	SV	SV
hoat4tian2	N	V
,	PUNC	PUNC
a1si7	CCL	CCL

	HunPos	me
,	PUNC	PUNC
『	PUNC	PUNC
sio2kiat4	ADV	N
』	V	PUNC
koh4	ADV	ADV
toa7liong5	V	SV
『	PUNC	PUNC
ko3iong7	ADV	V
』	V	PUNC
tong4te7	N	N
lang5	N	N
,	PUNC	PUNC

Tagging difficulty
Difficult for manual labor & HunPos

	HunPos	me
hoa1jin5	N	N
e5	AM	AM
long5bin5	N	N
ju2	ADV	SV
lai5	DL	V
ju2	V	SV
che7	N	SV
,	PUNC	PUNC

ju2 lai5 ju2 che7
? come ? plenty
"The more…the …er"

	HunPos	me
i2	V	V
tit4	N	V
tioh8	ADV	PVMOD
mo7ik8	V	N
siong7	N	PREP
e5	AM	AM
li7sun5	N	N
ui7	N	V
bok8tek8	N	N

i2 tit4 tioh8…
? get ?…
'In order to get…'

	HunPos	me
,	PUNC	PUNC
ui7	N	V
tioh8	ADV	PVMOD
beh4	AV	AV
seng4san3	V	V
kam4chia3	N	N
,	PUNC	PUNC

ui7 tioh8…
for ?…
'in order to…'

Summary

(1) Overall: Spoken character and written character had higher tagging accuracy; spoken romanization and written romanziation had lower tagging accuracy.

(2) The larger the training data, the higher the accuracy.

Summary

(3)The percentage of unseen words had an impact on the tagging accuracy.

(4) The number of model tags or tokens in the training data didn’t play a significant role in tagging accuracy. Certain tags got mixed up due to similar syntactic patterns.

References

Embree, Bernard L.M. 1984. A Dictionary of Southern Min. Taipei Language Institute.

Feng, Zhiwei. 2006. *Evolution and present situation of corpus research in China*. International Journal of Corpus Linguistics, 11:2, pp. 173-207. John Benjamins Publishing Company.

Halácsy, P., Kornai, A., and Oravecz, Cs. 2007. *HunPos - an open source trigram tagger*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume, Proceedings of the Demo and Poster Sessions, pp. 209-212. Prague, Czech Republic. 2007. Association for Computational Linguistics.

McEnery, Anthony and Xiao Zhonghua. 2004. *The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study*. Proceedings of the Fourth International Conference on Language Resources and Evalation (LREC) 2004, pp.1175-78. Available from: http://www.lancs.ac.uk/postgrad/xiaoz/papers/231.pdf

Zhang Ruixiong. 2009. *Huang Jia San Dai Mushi* (The three generations of priests). http://taigu.fhl.net/index.html.

Thank you.