Corpus-Linguistic and Cognitive Approaches to Determiner Usage in Chinese Student Writing

Testing the Fluctuation Hypothesis

Sven Albrecht

TU Chemnitz

05.02.2015

TECHNISCHE UNIVERSITÄT CHEMNITZ

Outline

1. Theory
2. Methodology
3. Analysis
4. Results
5. Discussion & Limitations
6. References

Outline: Theory

1. Theory
   - Definiteness & Specificity
   - Fluctuation Hypothesis
   - Missing Surface Inflection Hypothesis
   - Definiteness in Mandarin Chinese
   - Previous Studies
   - Research Questions
2. Methodology
3. Analysis
4. Results
5. Discussion & Limitations
6. References

Definition: Definiteness & Specificity

Informal definitions

(1) If a Determiner Phrase (DP) or the form [D NP] is …

a. [+definite], then the speaker and the hearer presuppose the existence of a unique individual in the set denoted by the NP.

b. [+specific], then the speaker intends to refer to a unique individual in the set denoted by the NP and considers this individual to possess some noteworthy property.

(from Ionin et al., 2004, p. 5)

Examples: Definiteness & Specificity

(2) a. Joan wants to present the prize to *the* winner
b. …but he doesn't want to receive it from her. (specific)
c. …so she'll have to wait around till the race finishes. (non-specific)
(from Lyons, 1999, p. 167, example (19))

(3) a. Peter intends to marry *a* merchant banker
b. …even though he doesn't get on at all with her. (specific)
c. …though he hasn't met one yet. (non-specific)
(from Lyons, 1999, p. 167, example (18))

Scope

(4) a. John met a stranger.
b. $\exists x\,(stranger(x)\ \&\ met(John, x))$

(from Lyons, 1999, p. 169, example (32))

(5) a. John didn’t meet a stranger.
b. $\sim \exists x\,(stranger(x)\ \&\ met(John, x))$
c. $\exists x\,(stranger(x)\ \&\ \sim met(John, x))$

(from Lyons, 1999, p. 169, example (33))

Example for (5-b):

(6) John didn’t meet a stranger; he couldn’t have, he knows everybody.

(from Lyons, 1999, p. 169, example (34))

Example for (5-c):

(7) John didn’t meet any stranger; he didn’t meet anyone.

(based on Lyons, 1999, p. 169)

Partitivity

(8) If a DP is [+partitive], it denotes an individual that is a member of a set introduced by previous discourse (c.f. Diesing, 1992; Enç, 1991).
(from Ionin et al., 2009, p. 14)

(9) [+partitive: explicit partitive]
Robert: He [Aaron] went to our local pet shop. This pet shop had five puppies and seven kittens, and Aaron loved all of them. But he could get only one! [...] Well, it was difficult for him to make up his mind. But finally, he got (a, the, –) puppy. Aaron went home really happy!

(10) [+partitive: implicit partitive]
Jane: Your friend Lucy looks really excited. What's going on?
Mary: She went to the airport to see her mother off, and ran into the Boston Red Sox team. She was very lucky - she got an autograph from (a, the, –) player.
(from Ionin et al., 2009, examples 23-24)

The Fluctuation Hypothesis

Definition

The Fluctuation Hypothesis (FH) for L2 article choice:

1. L2-learners have full access to the features that can underlie article choice cross-linguistically: the features [+definite] and [+specific].
2. L2-learners fluctuate between dividing English articles on the basis of definiteness vs. specificity, until the input leads them to choose the definiteness option.

(from Ionin et al., 2004, p. 8)

	[+definite] (target: the)	[-definite] (target: a)
[+specific]	correct use of *the*	**overuse of *the***
[-definite]	**overuse of *a***	correct use of *a*

Table 1: Predictions for Article Choice in Chinese L2 English (from Ionin et al. (2004))

The Missing Surface Inflection Hypothesis

Definition

[t]he Missing Surface Inflection Hypothesis (MSIH) proposes that L2 learners have unconscious knowledge of the functional projections and features underlying tense and agreement. However, learners sometimes have a problem with realization of surface morphology, such that they resort to non-finite forms [...]. (Prévost and White, 2000, p. 103)

- Full Transfer/Full Access: L1 final state = L2 initial state (Schwartz and Sprouse, 1996)
- issues in L2 production stem from morpholexical aspects rather than systematic syntactic deficits (Haznedar, 2001, p. 280)
- knowledge about the underlying abstract concepts of definiteness and the count/mass distinction is present in the learners' L1 and thus available in the L2 (Bergeron-Matoba, 2007)

Definiteness in Mandarin Chinese

- Mandarin Chinese lacks articles (Snape (2009), cited in Barrett and Chen (2011))
- identifiability marked by lexical, morphological and positional linguistic devices (Chen, 2004, p. 1151)
- lexical definiteness markers (Chen, 2004, p. 1151):
  - demonstratives
  - possessives
  - universal quantifiers
- positional (Chen, 2004, p. 1176)

Recent Development

Grammaticalization of demonstratives *zhe/na* and numeral *yi + CL* in definite and indefinite articles. (see Chen, 2004 for further references)

Definite Article

(11) SITUATIONAL:
Qing ba zhe/na zhang yizi ban dao na jian fangjian qu.
Please BA this/that CL chair move to that CL room go
Please move this/that chair to that room.

(12) DISCOURSE DEICTIC:
Ta xiang huiqu? Zhe ni ke buneng daying.
He want return this you surely cannot agree
He wants to go back? You surely cannot give your permission to that.

(from Chen, 2004, examples (49))

Indefinite Article

five stage grammaticalization process (Heine (1997) quoted in Chen (2003, p. 1170)):

(13) a. NUMERAL
I need *an hour* and *a half*.

b. PRESENTATIVE USE
A man came up the front stairs.

c. NON-IDENTIFIABLE SPECIFIC REFERENCE
He bought *a house* last year.

d. NON-IDENTIFIABLE NON-SPECIFIC REFERENCE
He wants to buy *a house* in this area; any house will do.

e. NON-REFERENTIAL USE
He is *a good chef*.

(from Chen, 2003, example (3))

Previous Studies

- Snape et al. (2006)
- White (2008)
- Ionin et al. (2009)

Previous Studies

- Snape et al. (2006)
  - Chinese, Japanese and Spanish L2ers
  - Japanese fluctuate as predicted by Fluctuation Hypothesis
  - Chinese fluctuate less than predicted
  - Spanish do not fluctuate (as predicted)
  - overuse of *a* in [+definite] [+specific] contexts
  - individual patterns concealed in group patterns (Hawkins et al., 2006)
  - difficulties matching features of vocabulary to underlying terminal nodes
  - Chinese pattern due to process of grammaticalization

Previous Studies

- Snape et al. (2006)
  - Chinese, Japanese and Spanish L2ers
  - Japanese fluctuate as predicted by Fluctuation Hypothesis
  - Chinese fluctuate less than predicted
  - Spanish do not fluctuate (as predicted)
  - overuse of *a* in [+definite] [+specific] contexts
  - individual patterns concealed in group patterns (Hawkins et al., 2006)
  - difficulties matching features of vocabulary to underlying terminal nodes
  - Chinese pattern due to process of grammaticalization

Previous Studies

- Snape et al. (2006)
- White (2008)
  - picture based elicited production task
  - 15 Chinese L2ers
  - clear *the* for *a* substitution pattern
  - *a* for *the* substitution almost unnoticeable → methodology
  - fluctuation patterns as predicted by Fluctuation Hypothesis

Previous Studies

- Snape et al. (2006)
- White (2008)
  - picture based elicited production task
  - 15 Chinese L2ers
  - clear *the* for *a* substitution pattern
  - *a* for *the* substitution almost unnoticeable → methodology
  - fluctuation patterns as predicted by Fluctuation Hypothesis

Previous Studies

- Snape et al. (2006)
- White (2008)
- Ionin et al. (2009)
    - 40 adult L1-Korean & 30 adult L1-Russian participants
    - written elicitation task & supplementary written narrative task
    - errors in [+definite] [-specific] & [-definite] [+specific] contexts
    - significant influence of definiteness & specificity (repeated-measures ANOVA)
    - 20 adult L1-Korean participants
    - significant overuse of *the* with implicit and explicit partitive indefinites (ANOVA)
    - learners associate *the* with [+specific] & [+partitive]

Previous Studies

- Snape et al. (2006)
- White (2008)
- Ionin et al. (2009)
    - 40 adult L1-Korean & 30 adult L1-Russian participants
    - written elicitation task & supplementary written narrative task
    - errors in [+definite] [-specific] & [-definite] [+specific] contexts
    - significant influence of definiteness & specificity (repeated-measures ANOVA)
    - 20 adult L1-Korean participants
    - significant overuse of *the* with implicit and explicit partitive indefinites (ANOVA)
    - learners associate *the* with [+specific] & [+partitive]

Previous Studies

- Snape et al. (2006)
- White (2008)
- Ionin et al. (2009)

Summary

Fluctuation Hypothesis confirmed in all three studies.
However: Chinese learners behave differently!

Previous Studies

- Snape et al. (2006)
- White (2008)
- Ionin et al. (2009)

Summary

Fluctuation Hypothesis confirmed in all three studies.
However: Chinese learners behave differently!

Issue

hardly any analytical statistics and no effect size reported

Research Questions

1. Do fluctuation patterns exist in Chinese learners’ article choice?
   1.1 How do semantic factors (partitivity and specificity) influence article choice?
   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
2. In which contexts does article misuse occur in the corpus data?
   2.1 Does lexical difficulty increase article misuse?
   2.2 How frequent is article misuse/does fluctuation occur in the corpus data?
3. Can similar fluctuation patterns be found in both data sets?

Outline: Methodology

1. Theory
2. Methodology
    - Grammaticality Judgment Test
    - Corpus
3. Analysis
4. Results
5. Discussion & Limitations
6. References

Online Survey

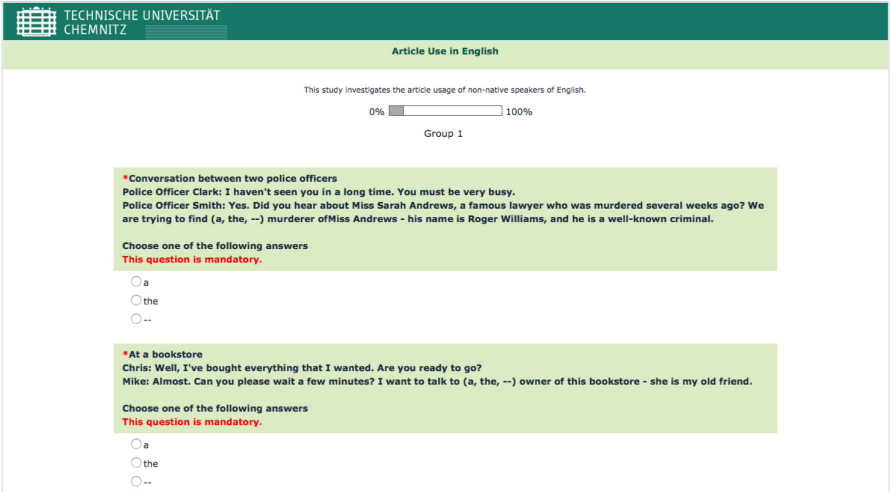
- LimeSurvey Software (Schmitz and LimeSurvey Project Team, 2015) provided by BPS Bildungsportal Sachsen GmbH[1]
- participant background information
  - year of birth, gender, L1 (multilingualism)
  - education (degree, major)
  - stays abroad, writing instruction (incl. teacher background)
  - questionnaire behavior
- forced choice elicitation test
  - simple definites
  - complex definites
  - partitivity
  - lexically difficult sentences

Major, unanticipated issue

participants unable to understand *multilingual*, despite additional explanation

[1] https://bildungsportal.sachsen.de/survey/index.php

Methodology
Grammaticality Judgment Test



TECHNISCHE UNIVERSITÄT CHEMNITZ

Article Use in English

This study investigates the article usage of non-native speakers of English.

0% 100%

Group 1

***Conversation between two police officers**
Police Officer Clark: I haven't seen you in a long time. You must be very busy.
Police Officer Smith: Yes. Did you hear about Miss Sarah Andrews, a famous lawyer who was murdered several weeks ago? We are trying to find (a, the, --) murderer ofMiss Andrews - his name is Roger Williams, and he is a well-known criminal.

Choose one of the following answers
This question is mandatory.

○ a
○ the
○ --

***At a bookstore**
Chris: Well, I've bought everything that I wanted. Are you ready to go?
Mike: Almost. Can you please wait a few minutes? I want to talk to (a, the, --) owner of this bookstore - she is my old friend.

Choose one of the following answers
This question is mandatory.

○ a
○ the
○ --

Figure 1: Online Survey (Forced Choice Elicitation Task)

Sources of Questions

53 questions in 7 groups, adopted from:

- Simple & Complex Definites (Ionin et al., 2004)
- Partitivity (Ko et al., 2010)
- Lexical Difficulty (Shifman et al., 1979; Sjöstrand, 1994)

Questions were reproduced unmodified, as in

Chrabaszcz and Jiang (2014), Crosthwaite (2014), DİKİLİTAŞ and Altay (2011), Hawkins et al. (2006), Ionin et al. (2009), Ionin and Montrul (2010), Montrul and Ionin (2012), Schönenberger (2014), Snape et al. (2006), and White (2008).

Questionnaire Examples: Lexical Difficulty

(14) It is in the nature of a program of this kind never to be finished, at least as long as it is of importance for (a, the, –) high-energy physics experimental community.

(adapted from Sjöstrand, 1994)

(15) Within the framework of (a, the, –) factorization hypothesis (i.e., vacuum insertions in all the channels, see sect. 6), the vacuum expectation values of these operators are connected with each other: [...]

(adapted from Shifman et al., 1979)

Questionnaire Examples: Scope & Speaker Knowledge

(16) At a bookstore
Chris: Well, I've bought everything that I wanted. Are you ready to go?
Mike: Almost. Can you please wait a few minutes? I want to talk to (a, the,–) owner of this bookstore - she is my old friend.

(17) Paul: Do you have time for lunch?
Sheila: No, I'm very busy. I am meeting with (a, the,–) president of our university, Dr. McKinley; it's an important meeting.

(18) In a clothing store
Clerk: May I help you?
Customer: Yes, please! I've rummaged through every stall, without any success. I am looking for (a, the,–) warm hat. It's getting rather cold outside.

(19) Rose: Let's go out to dinner with your brother Samuel tonight.
Alex: No, he is busy. He is having dinner with (a, the, –) manager of his office; I don't know who that is, but I'm sure that Samuel can't cancel this dinner.

(from Ionin et al., 2004, Appendix B)

Controlling Specificity, Scope & Speaker Knowledge

Specificity	Scope	Speaker Knowledge	Explanation
+specific	wide	yes	possible, c.f. (16)
+specific	wide	no	+specific = speaker knowledge
+specific	narrow	yes	+specific = wide scope
+specific	narrow	no	+specific = wide scope
+specific	no	yes	possible, c.f. (17)
+specific	no	no	+specific = speaker knowledge
-specific	wide	yes	-specific = narrow scope
-specific	wide	no	-specific = narrow scope
-specific	narrow	yes	-specific = no speaker knowledge
-specific	narrow	no	possible, c.f. (18)
-specific	no	yes	-specific = no speaker knowledge
-specific	no	no	possible, c.f. (19)

Table 2: Possible Combinations of Specificity, Scope & Speaker Knowledge

The Corpus Data

SYSU-Corpus	average length	Texts	Words
Master Theses (Linguistics)	16 900	25	422 535
Master Term Paper (Linguistics)	2500	86	216 278
Master Term Paper (FL Teaching)	2800	71	200 237
Bachelor Theses (Linguistics)	11 100	2	22 191
Bachelor Papers (Linguistics)	3000	2	5933
Total		186	867 174

Table 3: The *SYSU*-Corpus, from Küchler (2015, p. 105)

Issues:

- stratification, esp. gender
- ”dirty data”
- not part-of-speech tagged

Further Steps

POS Tagging

Stanford Log-linear Part-of-Speech Tagger[2] (as described in Toutanova, Klein, et al., 2003; Toutanova and Manning, 2000)

(20) Meanwhile_RB ,_, the_DT stack_VBP of_IN parallelism_NN enhances_VBZ language_NN force_NN ._.

Cleaning the Data

removing sentences that:

- start with a digit (headings)
- contain two or more consecutive whitespace characters (remnants of converted captions, graphs or tables)
- contain two or more consecutive non-alphanumeric character (encoding errors/general errors)

[2] available at http://nlp.stanford.edu/software/tagger.shtml

Further Steps: Random Sampling

- self-developed Python script
- sentence tokenization, pre-trained with Reuters Corpus (Natural Language Processing Toolkit (NLTK), Bird et al., 2009)
- Reservoir sampling algorithm by Alan G. Waterman as described in Knuth (1981)

```
def random_sampler(data, k):
sample = []
for n, line in enumerate(data):
 if n < k:
  sample.append(line.rstrip())
 else:
  r = random.randint(0, n)
  if r < k:
   sample[r] = line.rstrip()
return sample
```

Listing 1: Python Code

Further Steps: Sampling Omission Errors

Tag	Description
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural

Table 4: Relevant Tags from the Penn Treebank Tag Set (Santorini, 1990)

```
(VB|VBD|VBZ|VBN|VBG|VBP)\b\s\w+(NN|NNP|NNPS|NNS)
```

Listing 2: Regular Expression

Outline: Analysis

1. Theory

2. Methodology

3. Analysis
   - Expectations
   - Survey
   - Survey: Complex Definites
   - Survey: Partitivity, Simple Definites & Advanced Lexis
   - Corpus
   - Corpus: Error Examples

4. Results

5. Discussion & Limitations

6. References

Predicted Usage Patterns

	[+definite] (target: *the*)	[-definite] (target: *a*)
[+specific]	correct use of *the*	**overuse of *the***
[-definite]	**overuse of *a***	correct use of *a*

Table 5: Predictions for Article Choice in Chinese L2 English (adapted from Snape et al. (2006))

Survey Demographics

Male	Female
19	20

(a) Gender

Engineering	Humanities
22	17

(b) Discipline

Chinese	Not Chinese
37	2

(c) L1

Table 6: Survey Demographics I

Abroad	Not Abroad
6	33

(a) Studied Abroad

Received	Not Received
19	20

(b) Writing Instruction

Table 7: Survey Demographics II

Results: Total

	[+definite] (target *the*)		[-definite] (target *a*)	
L1 Chinese (n=37)	the	a	the	a
[+specific]	84%	14%	**10%**	90%
[-specific]	89%	**9%**	4%	93%

Table 8: Total Article Choice

	[+definite] (target *the*)	[-definite] (target *a*)
L1 Chinese (n=37)		
[+specific]	2%	0%
[-specific]	2%	3%

Table 9: Total Article Omission

Article Usage by Field

	[+definite] (target *the*)		[-definite] (target *a*)	
Engineering (n=22)	the	a	the	a
[+specific]	81%	17%	**14%**	86%
[-specific]	88%	**10%**	6%	93%

(a) Engineering

	[+definite] (target *the*)		[-definite] (target *a*)	
Humanities (n=15)	the	a	the	a
[+specific]	87%	8%	**4%**	95%
[-specific]	91%	**8%**	1%	94%

(b) Humanities

Table 10: Article Choice by Field

Omission by Field

	[+definite] (target *the*)	[-definite] (target *a*)
Engineering (n=22)		
[+specific]	1%	0%
[-specific]	2%	3%

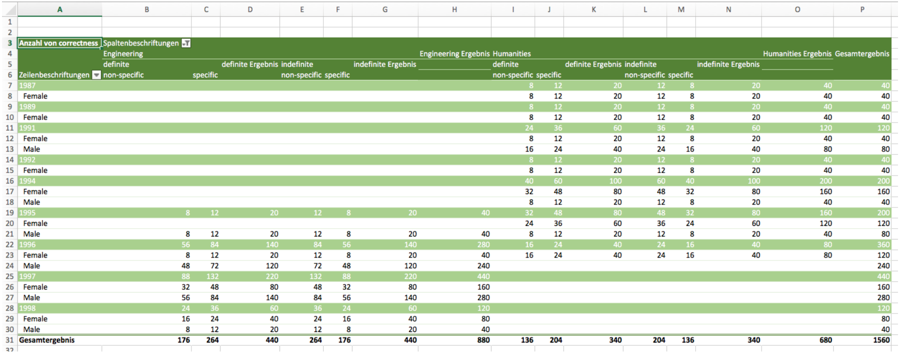
(a) Engineering

	[+definite] (target *the*)	[-definite] (target *a*)
Humanities (n=15)		
[+specific]	4%	1%
[-specific]	2%	5%

(b) Humanities

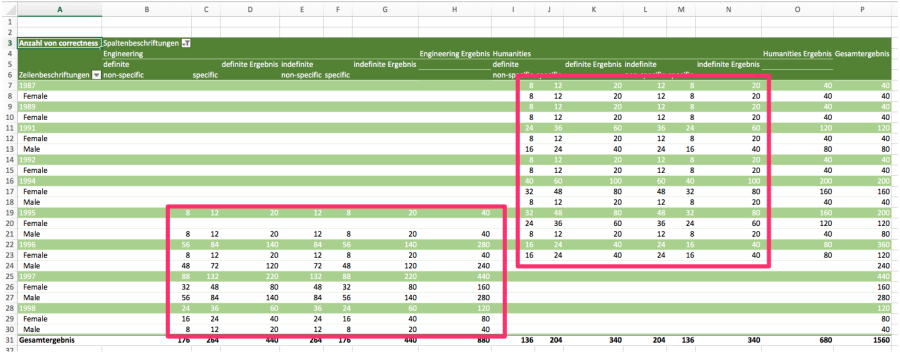
Table 11: Article Omission by Field

Data Cross-Tabulation



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3	Anzahl von correctness	Spaltenbeschriftungen														
4		Engineering						Engineering Ergebnis	Humanities						Humanities Ergebnis	Gesamtergebnis
5		definite		definite Ergebnis	indefinite		indefinite Ergebnis		definite		definite Ergebnis	indefinite		indefinite Ergebnis		
6	Zeilenbeschriftungen	non-specific	specific		non-specific	specific			non-specific	specific		non-specific	specific			
7	1987								8	12	20	12	8	20	40	40
8	Female								8	12	20	12	8	20	40	40
9	1989								8	12	20	12	8	20	40	40
10	Female								8	12	20	12	8	20	40	40
11	1991								24	36	60	36	24	60	120	120
12	Female								8	12	20	12	8	20	40	40
13	Male								16	24	40	24	16	40	80	80
14	1992								8	12	20	12	8	20	40	40
15	Female								8	12	20	12	8	20	40	40
16	1994								40	60	100	60	40	100	200	200
17	Female								32	48	80	48	32	80	160	160
18	Male								8	12	20	12	8	20	40	40
19	1995	8	12	20	12	8	20	40	32	48	80	48	32	80	160	200
20	Female								24	36	60	36	24	60	120	120
21	Male	8	12	20	12	8	20	40	8	12	20	12	8	20	40	80
22	1996	56	84	140	84	56	140	280	16	24	40	24	16	40	80	360
23	Female	8	12	20	12	8	20	40	16	24	40	24	16	40	80	120
24	Male	48	72	120	72	48	120	240								240
25	1997	88	132	220	132	88	220	440								440
26	Female	32	48	80	48	32	80	160								160
27	Male	56	84	140	84	56	140	280								280
28	1998	24	36	60	36	24	60	120								120
29	Female	16	24	40	24	16	40	80								80
30	Male	8	12	20	12	8	20	40								40
31	**Gesamtergebnis**	**176**	**264**	**440**	**264**	**176**	**440**	**880**	**136**	**204**	**340**	**204**	**136**	**340**	**680**	**1560**

Data Cross-Tabulation



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3	Anzahl von correctness	Spaltenbeschriftungen														
4		Engineering						Engineering Ergebnis	Humanities						Humanities Ergebnis	Gesamtergebnis
5		definite		definite Ergebnis	indefinite		indefinite Ergebnis		definite		definite Ergebnis	indefinite		indefinite Ergebnis		
6	Zeilenbeschriftungen	non-specific	specific		non-specific	specific			non-specific	specific		non-specific	specific			
7	1987								8	12	20	12	8	20	40	40
8	Female								8	12	20	12	8	20	40	40
9	1989								8	12	20	12	8	20	40	40
10	Female								8	12	20	12	8	20	40	40
11	1991								24	36	60	36	24	60	120	120
12	Female								8	12	20	12	8	20	40	40
13	Male								16	24	40	24	16	40	80	80
14	1992								8	12	20	12	8	20	40	40
15	Female								8	12	20	12	8	20	40	40
16	1994								40	60	100	60	40	100	200	200
17	Female								32	48	80	48	32	80	160	160
18	Male								8	12	20	12	8	20	40	40
19	1995	8	12	20	12	8	20	40	32	48	80	48	32	80	160	200
20	Female								24	36	60	36	24	60	120	120
21	Male	8	12	20	12	8	20	40	8	12	20	12	8	20	40	80
22	1996	56	84	140	84	56	140	280	16	24	40	24	16	40	80	360
23	Female	8	12	20	12	8	20	40	16	24	40	24	16	40	80	120
24	Male	48	72	120	72	48	120	240								240
25	1997	88	132	220	132	88	220	440								440
26	Female	32	48	80	48	32	80	160								160
27	Male	56	84	140	84	56	140	280								280
28	1998	24	36	60	36	24	60	120								120
29	Female	16	24	40	24	16	40	80								80
30	Male	8	12	20	12	8	20	40								40
31	Gesamtergebnis	176	264	440	264	176	440	880	136	204	340	204	136	340	680	1560

One Level Linear Regression Model I

Variable	p-Value
specificity*scope*sp_kn	**0.0448**
gender	0.25
writing instruction	0.573
definiteness	0.901

Table 12: Regression Model p-Values for Engineering Dataset

MLM: one-level mixed effects model (speaker as a random intercept)

- data needs to be transposed from wide form into long form
- multiple benefits over RM-ANOVA
  - no homogeneity of variance
  - no sphericity
  - can handle randomly missing data
- good for nested data
- random intercept

Variation Explained

$R^2_{total} = 0.146$, $R^2_{fixed} = 0.083$ and $R^2_{random} = 0.065$, random intercept $\sigma = 0.501$

One Level Linear Regression Model II

factor	tokens	logodds	application value	centered factor weight
definiteness				
indefinite	270	0.019	0.107	0.505
definite	249	−0.019	0.100	0.405
specificity*scope*sp_kn				
specific*no*yes	118	0.571	0.161	0.639
non-specific*no*no	134	0.212	0.119	0.553
specific*wide*yes	134	−0.221	0.082	0.445
non-specific*narrow*no	133	−0.563	0.060	0.363

Table 13: Regression Model Dataset [Engineering, year of birth $\geq$1996, not abroad]

Model Statistics

$df = 8, intercept = -2.562, AIC = 345.963, n = 519$

Results: Total

	simple (in)definites	partitivity	difficult
correct	81%	78%	48%
substitution	18%	22%	40%
omission	1%	1%	13%

Table 14: Results for Simple (In)definites, Partitivity and Lexical Difficulty

Results: by Field

	simple (in)definites	partitivity	difficult
correct	74%	76%	44%
substitution	25%	24%	53%
omission	1%	1%	2%

(a) Engineering

	simple (in)definites	partitivity	difficult
correct	91%	80%	51%
substitution	8%	20%	20%
omission	2%	0%	29%

(b) Humanities

Table 15: Results for Simple (In)definites, Partitivity and Lexically Complex Contexts

Corpus Samples

Sample	Words
the	1719
a	1886
omission	1786

Table 16: Random Sample Number of Words

	the	*a*	*ø*
substitution	0	0	–
overuse	15	5	3
omission	5	0	6

Table 17: Corpus Analysis Random Samples *the*, *a* and Omission

Examples

the

(21) Then what is its magic code to ***the*** *success*?

a

(22) The theory of communicative competence gives ***a*** *central importance* to sociocultural factors and stresses the ability for language use.

ø

(23) I am so excited to have ***ø*** *opportunity* to learn Chinese in this school as an exchange student.

(Examples from *SYSU-C*, my emphasis)

Outline: Results

1. Theory

2. Methodology

3. Analysis

4. Results
   - Analysis Results
   - Answering the Research Questions

5. Discussion & Limitations

6. References

Results

General

- article errors occur in both datasets
- omission usually < 5%

Survey Data

- highest substitution rates in [+definite] [+specific] contexts (simple definites) (c.f. Snape et al., 2006)
- Engineering: 14% substitution in [-definite] [+specific] contexts
- 20-24% substitution for partitivity (c.f. Ionin et al., 2009)
- interaction group *specificity*, *scope* and *speaker knowledge*
- no influence of definiteness on variation in regression model

Corpus Data

- no fluctuation and no substitution errors in corpus samples
- overuse of (in)definite articles with generic, non-specific reference

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?
   ↪ no *fluctuation* patterns, but error patterns

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?
   ↪ no *fluctuation* patterns, but error patterns
   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?

   ↪ no *fluctuation* patterns, but error patterns

   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   ↪ see above

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?
   ↪ no *fluctuation* patterns, but error patterns

   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   ↪ see above

   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
   ↪ older humanities students > young engineering students, no influence of writing instruction

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?

   ↪ no *fluctuation* patterns, but error patterns

   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   ↪ see above

   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
   ↪ older humanities students > young engineering students, no influence of writing instruction

2. In which contexts does article misuse occur in the corpus data?

   ↪ mostly generic, non-specific reference

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?

   ↪ no *fluctuation* patterns, but error patterns

   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   ↪ see above

   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
   ↪ older humanities students > young engineering students, no influence of writing instruction

2. In which contexts does article misuse occur in the corpus data?

   ↪ mostly generic, non-specific reference

   2.1 Does lexical difficulty increase article misuse?
   ↪ cannot be answered with the data at hand

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?

   ↪ no *fluctuation* patterns, but error patterns

   1.1 How do semantic factors (partitivity and specificity) influence article
   ↪ *specificity*, *scope* and *speaker knowledge* interact

   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
   ↪ see above

   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
   ↪ older humanities students > young engineering students, no influence of writing instruction

2. In which contexts does article misuse occur in the corpus data?

   ↪ mostly generic, non-specific reference

   2.1 Does lexical difficulty increase article misuse?
   ↪ cannot be answered with the data at hand

   2.2 How frequent is article misuse/does fluctuation occur in the corpus data?
   ↪ misuse < 1%

Answering the Research Questions

1. Do fluctuation patterns exist in Chinese learners' article choice?
   ↪ no *fluctuation* patterns, but error patterns
   1.1 How do semantic factors (partitivity and specificity) influence article
       ↪ *specificity*, *scope* and *speaker knowledge* interact
   1.2 How do pragmatic factors (scope and speaker knowledge) influence article choice?
       ↪ see above
   1.3 How do sociolinguistic factors ([age], [gender], [major], [stays abroad], writing instruction) influence article choice?
       ↪ older humanities students > young engineering students, no influence of writing instruction
2. In which contexts does article misuse occur in the corpus data?
   ↪ mostly generic, non-specific reference
   2.1 Does lexical difficulty increase article misuse?
       ↪ cannot be answered with the data at hand
   2.2 How frequent is article misuse/does fluctuation occur in the corpus data?
       ↪ misuse < 1%
3. Can similar fluctuation patterns be found in both data sets?
   ↪ no *fluctuation* patterns, only equally low omission rates

Outline: Discussion & Limitations

1. Theory
2. Methodology
3. Analysis
4. Results
5. Discussion & Limitations
   - Discussion
   - Limitations
6. References

Discussion of Results

Why is there no fluctuation?

- grammaticalization process
- advanced learners

Omission Rates

- humanities higher omission rate = awareness of zero article?
- zero article is more difficult?
- humanities/philology = higher language awareness?

Corpus Sample Size

- similar number of words
- similar number of instances
- **but**: might be too small, quantitative analysis might be better

Limitations

General

- education differentiation suboptimal
- learners too advanced? (esp. in corpus data)

Statistics

- regression model only accounts for 8.3% of the variation
- statistic modeling needs more and better stratified data
- statistic modeling RM-ANOVA vs. MLM
- not enough data for humanities, partitivity & lexical complexity models

Cultural

- forced choice elicitation task $\approx$ multiple choice exam
- cultural reference caused confusion in forced choice elicitation task (*Boston Red Sox team*)

Outline: References

1. Theory
2. Methodology
3. Analysis
4. Results
5. Discussion & Limitations
6. References

References I

Barrett, N. E. & Chen, L.-m. (2011). English article errors in Taiwanese college students' EFL writing. *Computational Linguistics and Chinese Language Processing, 16*(3-4), 1–20.

Bergeron-Matoba, J. (2007). Acquisition of the English article system in SLA and the Missing Surface Inflection Hypothesis.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1st). O'Reilly Media, Inc.

Chen, P. (2003). Indefinite determiner introducing definite referent: a special use of 'yi 'one'+classifier' in Chinese. *Lingua, 113*(12), 1169–1184.

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics, 42*(6), 1129–1184.

Chrabaszcz, A. & Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: A cross-linguistic comparison. *Second Language Research, 30*(3), 351–379.

References II

Crosthwaite, P. R. (2014). Definite Discourse-New Reference in L1 and L2: A Study of Bridging in Mandarin, Korean, and English. *Language Learning, 64*(3), 456–492.

Diesing, M. (1992). *Indefinites*. Linguistic Inquiry Monographs. Cambridge.

DİKİLİTAŞ, K. & Altay, M. (2011). Acquisition sequence of four categories of non-generic use of the English definite article THE by Turkish speakers. *Novitas-ROYAL (research on Youth and Language), 5*(2), 183–198.

Enç, M. (1991). The semantics of specificity. *Linguistic inquiry.*

Hawkins, R., Al-Eid, S., Almahboob, I., Athanasopoulos, P., Chaengchenkit, R., Hu, J., … Jiang, A. (2006). Accounting for English article interpretation by L2 speakers. *EUROSLA yearbook, 6*(1), 7–25.

Haznedar, B. (2001). The acquisition of the IP system in child L2 English. *Studies in second language acquisition, 23*(01), 1–39.

Heine, B. (1997). *Cognitive Foundation of Grammar*. Oxford University Press.

Ionin, T., Ko, H., & Wexler, K. (2009). L2-acquisition of English articles by Korean speakers. *The handbook of East Asian psycholinguistics, 3*, 286–304.

References III

Ionin, T., Ko, H., & Wexler, K. (2004). Article Semantics in L2 Acquisition: The Role of Specificity. *Language Acquisition, 12*(1), 3–69.

Ionin, T. & Montrul, S. (2010). The Role of L1 Transfer in the Interpretation of Articles with Definite Plurals in L2 English. *Language Learning, 60*(4), 877–925.

Knuth, D. E. (1981). *The Art of Computer Programming* (2nd ed.). Seminumerical Algorithms. Reading, Massachusetts: Addison-Wesley.

Ko, H., Ionin, T., & Wexler, K. (2010). The role of presuppositionality in the second language acquisition of English articles. *Linguistic inquiry, 41*(2), 213–254.

Küchler, J. (2015). Usages of *may* and *will* in Chinese and German Student Writings. In J. Schmied (Ed.), *Academic writing for south eastern europe* (pp. 99–118). Göttingen: Cuvillier.

Lyons, C. (1999). *Definiteness* . (1. publ.). Cambridge [u.a.]: Cambridge Univ. Press.

Montrul, S. & Ionin, T. (2012). Dominant Language Transfer in Spanish Heritage Speakers and Second Language Learners in the Interpretation of Definite Articles. *The Modern Language Journal, 96*(1), 70–94.

References IV

Prévost, P. & White, L. (2000). Missing Surface Inflection or Impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research, 16*(2), 103–133.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision).

Schmitz, C. & LimeSurvey Project Team. (2015). LimeSurvey: An Open Source survey tool.

Schönenberger, M. (2014). Article use in L2 English by L1 Russian and L1 German speakers. *Zeitschrift für Sprachwissenschaft, 33*(1), 77–105.

Schwartz, B. D. & Sprouse, R. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research, 12*(1), 40–72.

Shifman, M. A., Vainshtein, A. I., & Zakharov, V. I. (1979). QCD and resonance physics. theoretical foundations. *Nuclear Physics B, 147*(5), 385–447.

Sjöstrand, T. (1994). High-energy-physics event generation with PYTHIA 5.7 and JETSET 7.4. *Computer Physics Communications, 82*(1), 74–89.

References V

Snape, N. (2009). Exploring Mandarin Chinese speakers' L2 article use. *Representational deficits in SLA, Studies in honor of Roger Hawkins*, 27–53.

Snape, N., Ting, H.-C., & Leung, Y.-k. I. (2006). Comparing Chinese, Japanese and Spanish Speakers in L2 English Article Acquisition: Evidence against the Fluctuation Hypothesis?

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *The 2003 conference of the north american chapter of the association for computational linguistics* (pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics.

Toutanova, K. & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 joint sigdat conference on empirical methods in natural language processing and very large corpora: held in conjunction with the 38th annual meeting of the association for computational linguistics-volume 13* (pp. 63–70). Association for Computational Linguistics.

References VI

White, L. (2008). Different? Yes. Fundamentally? No. Definiteness effects in the L2 English of Mandarin speakers. In *Proceedings of the 9th generative approaches to second language acquisition conference (gasla 2007)* (pp. 251–261).