Language Technology Lab Seminars@University of Cambridge

Corpus-Based Analysis of the Canonical Word Order of Double Object Constructions

Ryohei Sasano

2 Nov 2017

Topic: The word order of Japanese double object constructions

- This research mainly arose from a linguistic interest
- Possible word orders
  - Japanese has a much freer word order than English
  - In case of double object constructions, the position of the verb is fixed,
  - but the positions of its NOM, DAT, and ACC arguments can be scrambled
  - Word orders are different, but they have essentially the same meaning:

Ken showed a camera to Aya.

a.	ケンが	アヤに	カメラを	見せた。
	Ken-NOM	Aya-DAT	camera-ACC	showed
b.	ケンが	カメラを	アヤに	見せた。
	Ken-NOM	camera-ACC	Aya-DAT	showed
c.	アヤに	ケンが	カメラを	見せた。
	Aya-DAT	Ken-NOM	camera-ACC	showed
d.	アヤに	カメラを	ケンが	見せた。
	Aya-DAT	camera-ACC	Ken-NOM	showed
e.	カメラを	ケンが	アヤに	見せた。
	camera-ACC	Ken-NOM	Aya-DAT	showed
f.	カメラを	アヤに	ケンが	見せた。
	camera-ACC	Aya-DAT	Ken-NOM	showed

Contents

- Motivation of the study
- Japanese double object construction
- Corpus-based analysis
- Conclusion & Future directions

Background of the study

- Many studies on the canonical word order of Japanese double object constructions
  - Theoretical studies [Hoji'85; Miyagawa+'04]
  - Psychological experiment-based studies [Koizumi+'04; Nakamoto+'06; Takimoto+'15@LSJ]
  - Brain science studies [Koso+'04; Inubushi+'09]

- Most of them require either manual analyses or measurements of human characteristics
  - e.g. brain activities, reading times, etc.

Research Question: What findings can be derived from a corpus based approach?

Corpus-based approach

- Assumption:
  - there is a relation between the canonical word order and the proportion of each word order
- The proportion of each word order can be collected from a large corpus

言葉に 愛情を 感じる
word-DAT affection-ACC feel

DAT-ACC: (97.5%)

愛情を 言葉に 感じる
affection-ACC word-DAT feel

ACC-DAT: (2.5%)

(φ_I feel the affection in φ_{your} words.)

デートに 女性を 誘う
date-DAT woman-ACC ask

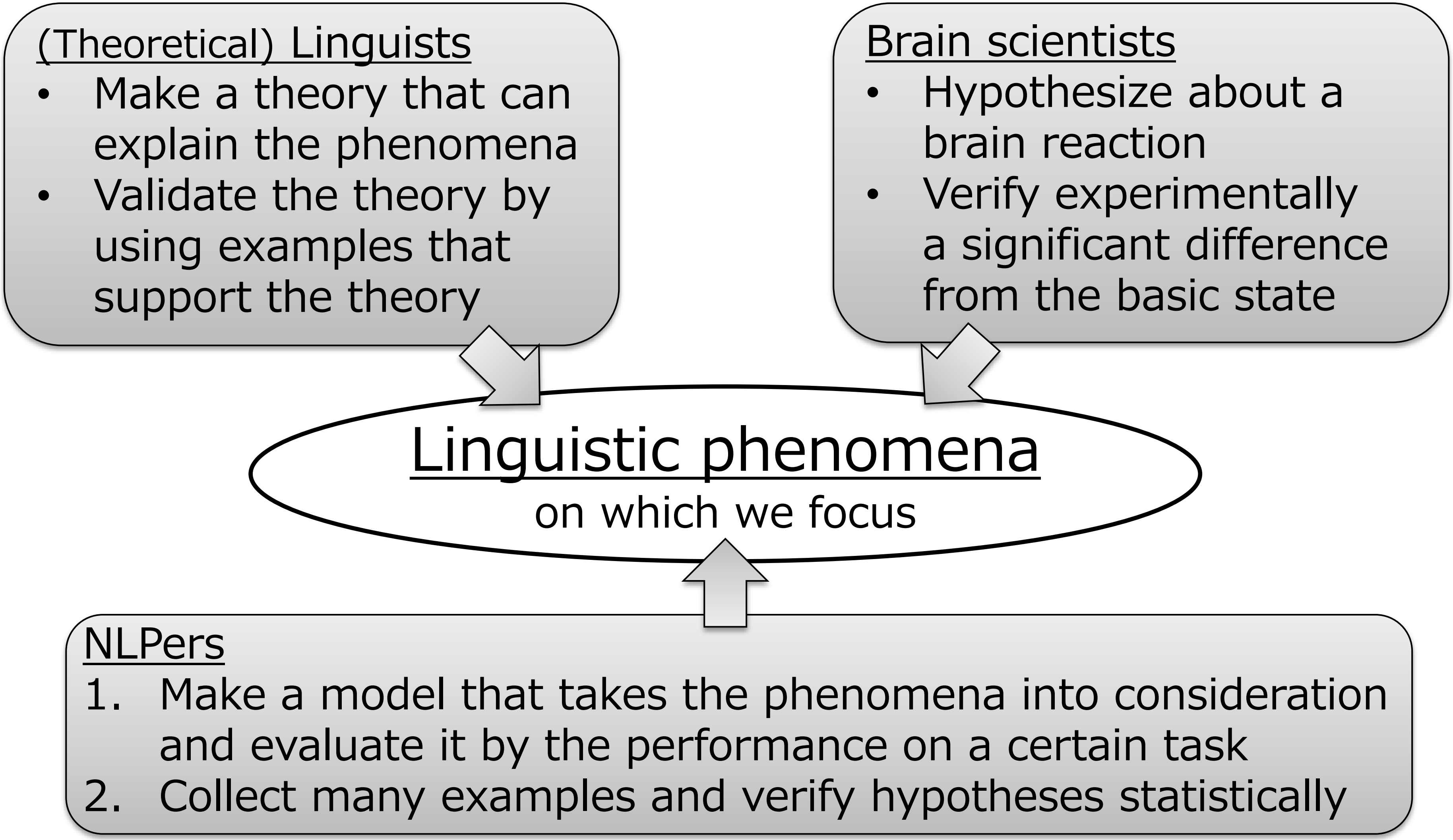
DAT-ACC: (0.4%)

女性を デートに 誘う
woman-ACC date-DAT ask

ACC-DAT: (99.6%)

(φ_I ask a woman out on a date.)

Approaches to linguistic phenomena



(Theoretical) Linguists

- Make a theory that can explain the phenomena
- Validate the theory by using examples that support the theory

Brain scientists

- Hypothesize about a brain reaction
- Verify experimentally a significant difference from the basic state

Linguistic phenomena

on which we focus

NLPers

1. Make a model that takes the phenomena into consideration and evaluate it by the performance on a certain task
2. Collect many examples and verify hypotheses statistically

Comparison of the approaches

	NLP/CL	Linguistics	Brain Science
Cost	🙂	😐	🙁
Scalability	🙂	🙁	🙁
Objectivity	🙂	😐	🙂
Reliability	🙁	🙂	😐
Immediacy	🙁	😐	🙂

Toward analysis of a phenomena that requires to take many combinations into account

1. Making a hypothesis on the phenomena using NLP techniques with a very large corpus
2. Verifying the hypothesis more precisely by using approaches based on linguistics or brain science

Contents

- Motivation of the study
- Japanese double object construction
- Corpus-based analysis
- Conclusion & Future directions

Relevant Japanese grammar

- Japanese word order is basically SOV, but it does not mark syntactic relations and is often scrambled
- Postpositional particles function as case markers
  - Nominative, dative, and accusative cases are represented by “が (ga)”, “に (ni)”, “を (wo)”, respectively
- Double object construction is a construction that contains two objects
  - In Japanese, they typically appear accompanying case particles “に” (dative) and “を” (accusative)
  - There are three arguments including the subject that accompanies “が” (nominative)

Six possible word orders

a. ケンが アヤに カメラを 見せた。 DAT-ACC
Ken-NOM Aya-DAT camera-ACC showed

b. ケンが カメラを アヤに 見せた。 ACC-DAT
Ken-NOM camera-ACC Aya-DAT showed

c. アヤに ケンが カメラを 見せた。
Aya-DAT Ken-NOM camera-ACC showed

d. アヤに カメラを ケンが 見せた。
Aya-DAT camera-ACC Ken-NOM showed

e. カメラを ケンが アヤに 見せた。
camera-ACC Ken-NOM Aya-DAT showed

f. カメラを アヤに ケンが 見せた。
camera-ACC Aya-DAT Ken-NOM showed

Related study

- Three major claims:
  1. [Hoji’85] argues the DAT-ACC order is canonical for all cases
  2. [Matsuoka’03] argues they have two canonical word orders, the DAT-ACC and ACC-DAT orders, depending on the verb types
  3. [Miyagawa’97] asserts that both the DAT-ACC and ACC-DAT orders are canonical for all cases
- Note that, the definition of the term *“canonical word order”* varies from study to study
- We basically adopt the definition and position:
  - The order that native Japanese speakers feel most natural
  - Only one canonical order for one tuple of a verb and arguments
  - The canonical word orders can be different for different tuples

Features related to word order

- A number of features that affect word order
  - Long arguments is placed far from the verb; short arguments is placed near the verb
  - Old, predictable information is placed first; new, unpredictable information is placed last

- We are interested in the canonical order
  - We do not take these features into account
  - We assume that these features can be ignored by using a very large corpus and analyzing based on the statistics

Claims to verify in this study

A) The DAT-ACC order is canonical [Hoji'85] No

B) There are two canonical word orders, the DAT-ACC and the ACC-DAT order, depending on the verb types [Matsuoka'03] No

C) An argument whose grammatical case is infrequently omitted with a given verb tends to be placed near the verb Yes

D) The canonical word order varies depending on the semantic role and animacy of the dative argument [Matsuoka'03] Yes

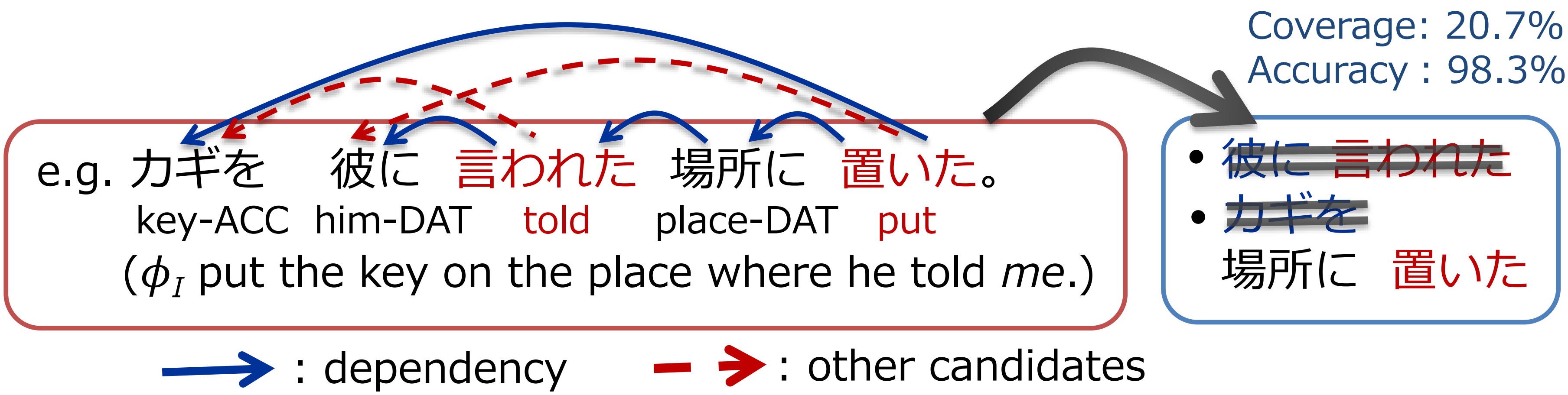
E) An argument that frequently co-occurs with the verb tends to be placed near the verb Yes

Contents

- Motivation of the study
- Japanese double object construction
- Corpus-based analysis
- Conclusion & Future directions

Example collection

- Difficulty:
  - Automatically collected examples sometimes include inappropriate ones
- Solution:
  - We extract examples from a corpus consisting of more than 10 billion Web sentences
  - We use only unambiguous parts of dependency parses, and collect the verb that had more than 500 different examples of dative and accusative argument pairs



Coverage: 20.7%
Accuracy : 98.3%

e.g. カギを　彼に　言われた　場所に　置いた。
key-ACC　him-DAT　told　place-DAT　put
(ϕ_I put the key on the place where he told *me*.)

- ~~彼に 言われた~~
- ~~カギを~~
  場所に　置いた

: dependency　　: other candidates

Statistics

- Corpus size:
  - > 10 billion unique sentences
- # of verbs that had 500 different examples:
  - 648 (all of which are ditransitive verbs)
- # of occurrences of each verb:
  - Average: 350k, Median: 83k
- # of extracted examples that include both dative and accusative arguments:
  - Average: 38k, Median: 9k

Word order for each verb

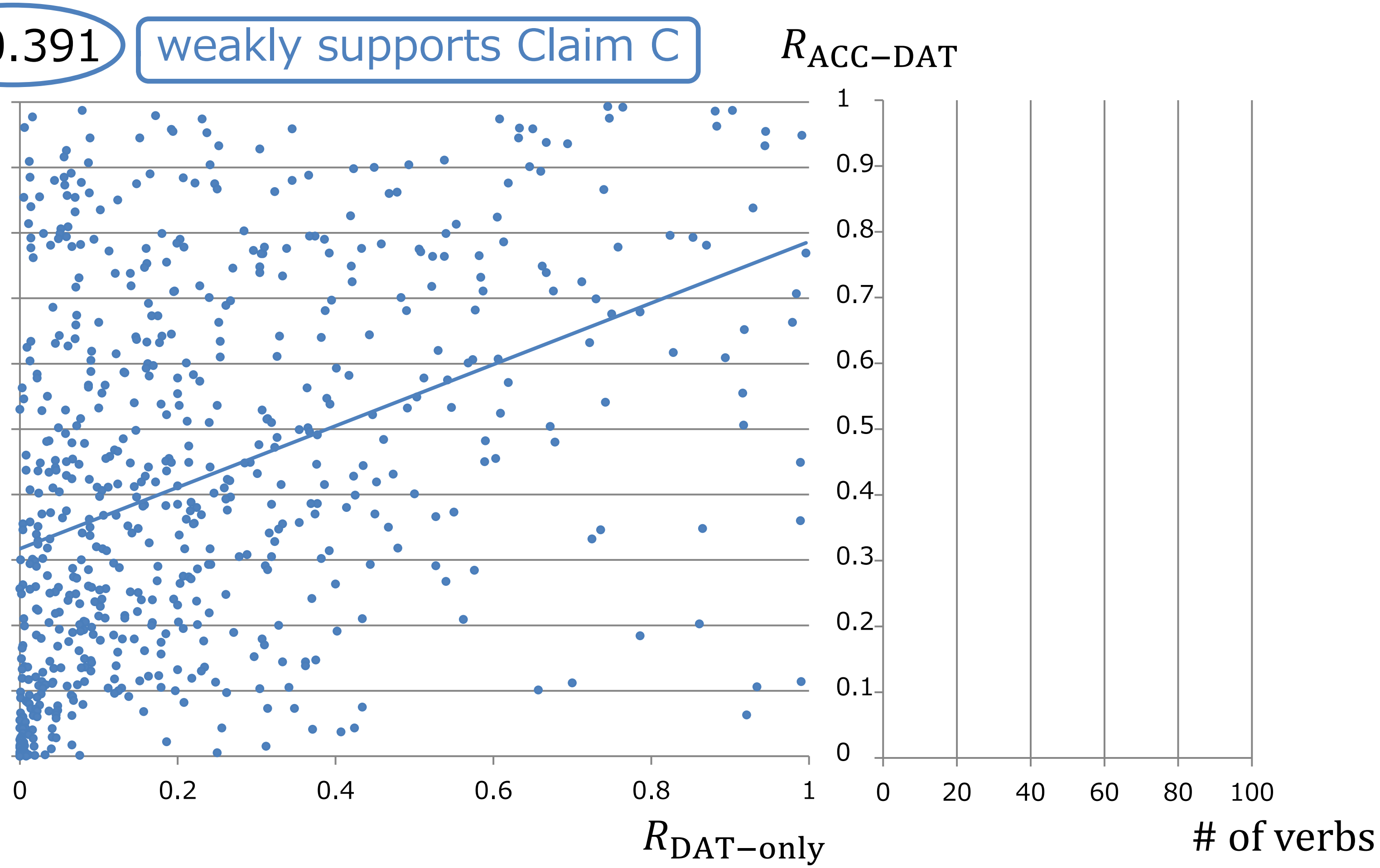
- Claims A and C:
  A) The DAT-ACC order is canonical [Hoji’85]
  C) An argument whose case is infrequently omitted with a given verb tends to be placed near the verb

女性を	デートに	誘う
woman-ACC	date-DAT	ask

- We examine the relation between
  - the proportion of the DAT only example $R_{\mathrm{DAT-only}}$
  - the proportion of the ACC-DAT order $R_{\mathrm{ACC-DAT}}$

$$R_{\mathrm{DAT-only}} = \frac{N_{\mathrm{DAT-only}}}{N_{\mathrm{DAT-only}} + N_{\mathrm{ACC-\ only}}}, \quad R_{\mathrm{ACC-DAT}} = \frac{N_{\mathrm{ACC-DAT}}}{N_{\mathrm{DAT-ACC}} + N_{\mathrm{ACC-DAT}}}$$

Relation between $R_{\mathrm{DAT-only}}$ and $R_{\mathrm{ACC-DAT}}$



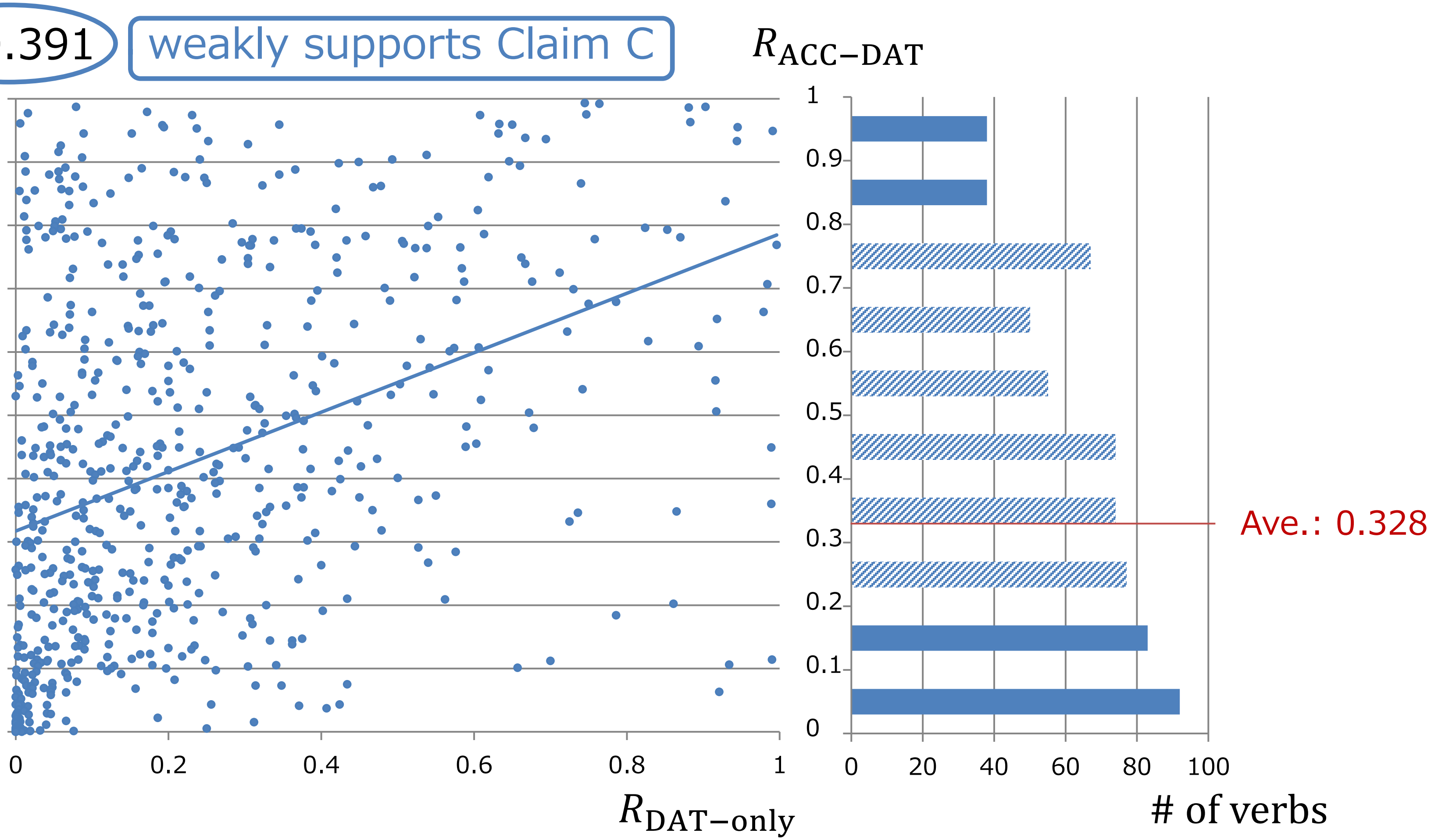
ρ: 0.391

weakly supports Claim C

Relation between $R_{\mathrm{DAT-only}}$ and $R_{\mathrm{ACC-DAT}}$

ρ: 0.391

weakly supports Claim C

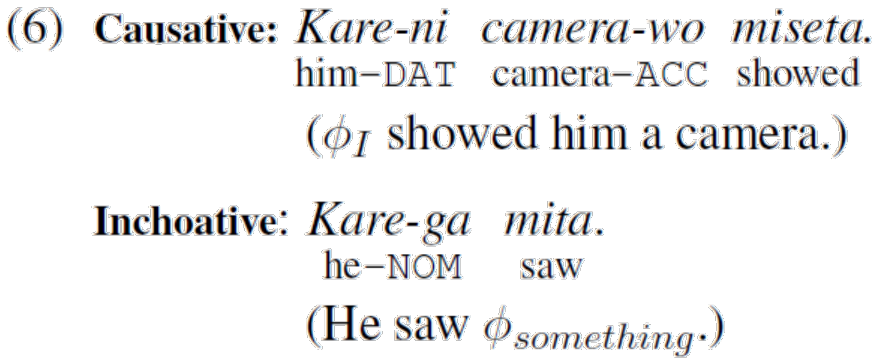


Ave.: 0.328

The preferred word order cannot be determined even if the verb is given in most cases ⇔ Claim A

Pass- and show- type

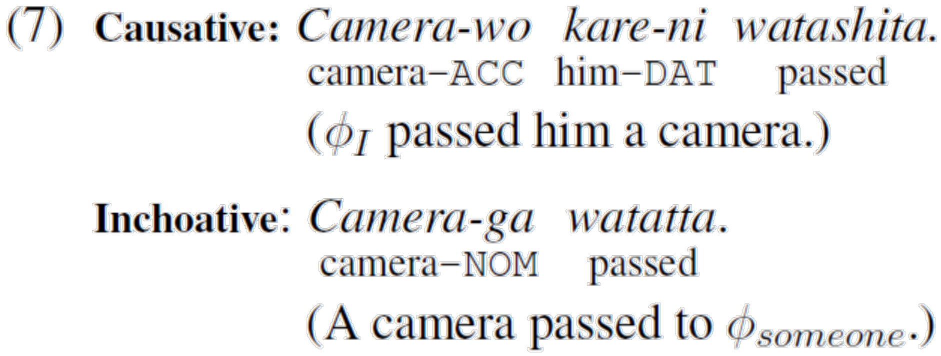
- Show-type: the dative argument is the subject of its corresponding inchoative sentence



(6) **Causative:** *Kare-ni camera-wo miseta.*
him−DAT camera−ACC showed
(ϕ_I showed him a camera.)

Inchoative: *Kare-ga mita.*
he−NOM saw
(He saw $\phi_{something}$.)

- Pass-type: the accusative argument is the subject of its corresponding inchoative sentence



(7) **Causative:** *Camera-wo kare-ni watashita.*
camera−ACC him−DAT passed
(ϕ_I passed him a camera.)

Inchoative: *Camera-ga watatta.*
camera−NOM passed
(A camera passed to $\phi_{someone}$.)

Word order and verb type

- Claim B: the DAT-ACC is canonical for show-type;
  the ACC-DAT is canonical for pass-type verbs
  - Classification based on causative-inchoative alternation

Show-type		Pass-type			
Verb	$R_{\mathrm{ACC-DAT}}$	verb	$R_{\mathrm{ACC-DAT}}$	Verb	$R_{\mathrm{ACC-DAT}}$
知らせる(notify)	0.522	戻す(put back)	0.771	落とす(drop)	0.351
預ける(deposit)	0.399	泊める(lodge)	0.748	漏らす(leak)	0.332
事付ける(request)	0.386	包む(wrap)	0.603	浮かべる(float)	0.255
悟す(adomish)	0.325	伝える(inform)	0.522	向ける(direct)	0.251
見せる(show)	0.301	載せる(place on)	0.496	残す(leave)	0.238
被せる(cover)	0.256	届ける(deliver)	0.491	埋める(bury)	0.223
教える(teach)	0.235	並べる(range)	0.481	混ぜる(blend)	0.200
授ける(give)	0.186	返す(give back)	0.448	当てる(hit)	0.185
浴びせる(shower)	0.177	ぶつける(knock)	0.436	掛ける(hang)	0.108
貸す(lend)	0.118	付ける(attach)	0.368	重ねる(pile)	0.084
着せる(dress)	0.113	渡す(pass)	0.362	建てる(build)	0.069
Macro average	0.274			Macro average	0.367

The difference is not significant and even in the case of pass-type verbs, the DAT-ACC order is dominant ⇔ Claim B

Word order and semantic role

- Claim D:
  - ACC-DAT is more preferred when the semantic role of the DAT is inanimate Goal than when the role is animate Possessor
- We collect the examples that satisfied:

A) ACC=ARTIFACT & DAT=**PLACE-INSTITUTION**

B) ACC=ARTIFACT & DAT=**PERSON**

本を　**学校**に　返した。
book-ACC **school**-DAT returned
(ϕ returned the book to school.)

先生に　本を　返した。
teacher-DAT book-ACC returned
(ϕ returned the book to the teacher.)

- Extract verbs that have at least 100 examples of both types
  - Out of 126 verbs, 64 verbs show the trend that **Type-A** prefers the ACC-DAT order more than **Type-B** does, and only 30 verbs have the opposite trend

supports Claim D

Word order for each tuple of a verb and its arguments

- Claim E:
  - An argument that frequently co-occurs with the verb tends to be placed near the verb

女性を	デートに	誘う
woman-ACC	date-DAT	ask

- We examined the relation between $R_{\mathrm{ACC-DAT}}$ and the degree of co-occurrence of a verb and its argument
  - We investigated 2302 tuples of a verb and its arguments that appear more than 500 times in the corpus
  - used the difference of NPMIs for measuring the degree of co-occurrence: $\mathrm{NPMI}(n_{\mathrm{DAT}}, v) - \mathrm{NPMI}(n_{\mathrm{ACC}}, v)$,

$$\text{where } \mathrm{NPMI}(n_c, v) = \frac{\mathrm{PMI}(n_c, v)}{-\log(p(n_c, v))}$$

a normalized version of PMI
The value ranges between [-1,+1]

Effects of idiomatic expression

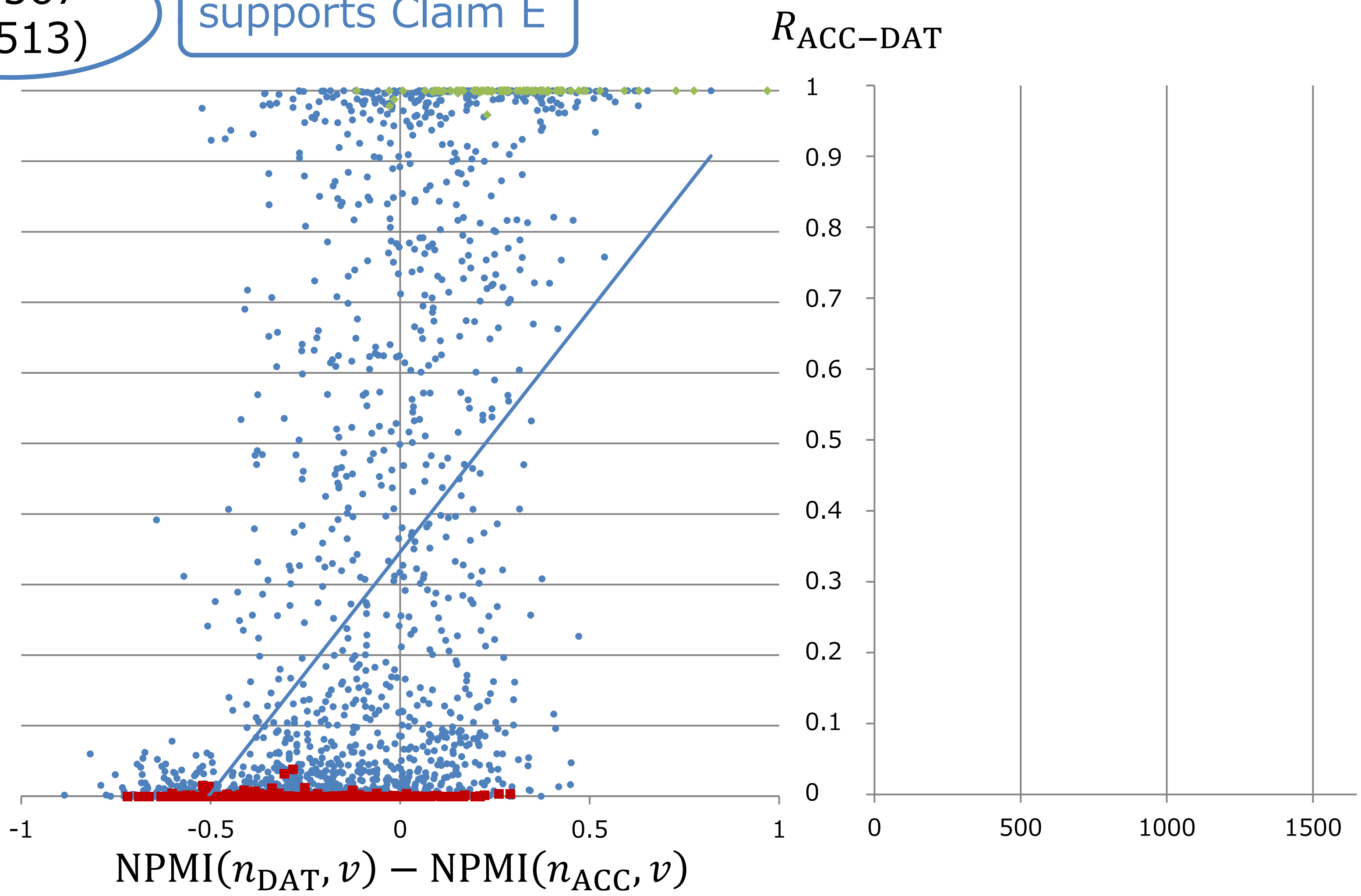
- One of the typical examples that satisfy Claim E is an idiomatic expression
  - A verb and its argument that are used as an idiom co-occur frequently and usually placed adjacent
  - We thus investigated to what extent idiomatic expressions affected the results

- We manually judged whether the verb and the adjacent argument are used as an idiom
  - Verb/ACC are judged as idiomatic for 404 out of 2302
  - Verb/DAT are judged as idiomatic for 84 out of 2302

Relation between $\mathrm{NPMI}(n_{\mathrm{DAT}}, v) - \mathrm{NPMI}(n_{\mathrm{ACC}}, v)$ and $R_{\mathrm{ACC-DAT}}$

ρ: 0.567 (0.513)

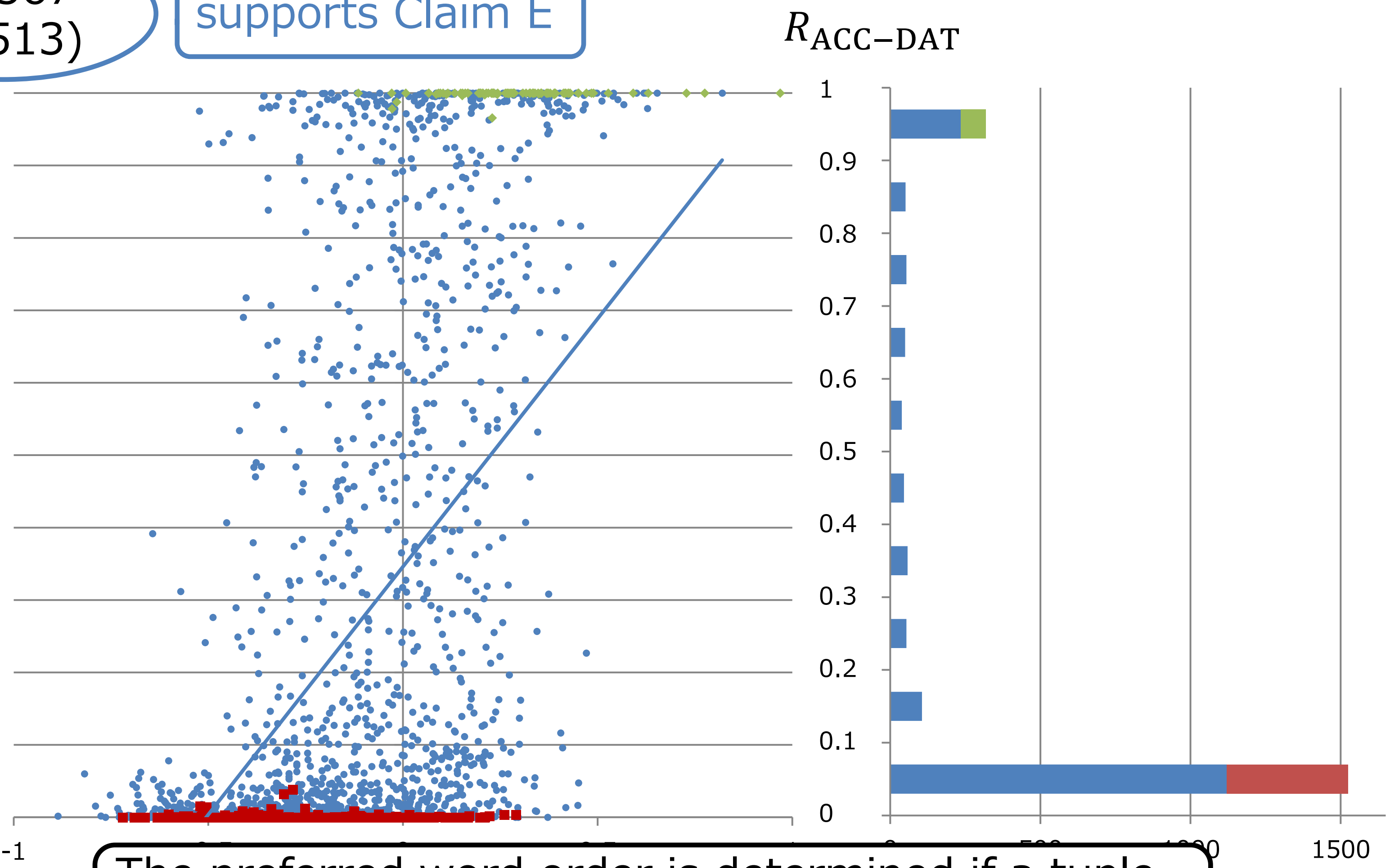
supports Claim E



Relation between $\mathrm{NPMI}(n_{\mathrm{DAT}}, v) - \mathrm{NPMI}(n_{\mathrm{ACC}}, v)$ and $R_{\mathrm{ACC-DAT}}$

ρ: 0.567
(0.513)

supports Claim E



The preferred word order is determined if a tuple of a verb and its arguments is given.

Conclusion: our analysis suggests

1. The canonical word order of Japanese double object constructions varies from verb to verb
2. There is only a weak relation between the canonical word order and the verb type
3. An argument whose grammatical case is infrequently omitted with a given verb tends to be placed near the verb
4. The canonical word order varies depending on the semantic role of the dative argument
5. An argument that frequently co-occurs with the verb tends to be placed near the verb

Future directions

1. Further verification of the findings depending on more reliable methodology such as brain science approach

2. Further investigation of the relation of semantic role and word order

3. Analysis of word order that takes context into consideration