A Radically Data-Driven Construction Grammar:

English Constructions of Letting and Vector Space Models

Natalia Levshina
F.R.S. – FNRS
Université catholique de Louvain

Constructions of Letting

The professor	let allowed permitted	her students	(to) use	their course notes

Constructions of Letting

The professor	let allowed permitted AUXILIARY	her students	(to) use	their course notes

Constructions of Letting

The professor	let allowed permitted	her students	(to) use EFFECTED PREDICATE (Vinf)	their course notes

Constructions of Letting

The professor	let allowed permitted	her students	(to) use	their course notes
CAUSER				

Constructions of Letting

The	let	her	(to) use	their
professor	allowed	students		course
	permitted			notes
		CAUSEE		

The aims of the study

Top-down approach

- pinpoint contextual factors identified on the basis of previous research (+intuition) that can account for the speaker’s choice between the cx
- fit a multivariate model to test if these variables help predict the use of cxs

The aims of the study

Top-down approach

- pinpoint contextual factors identified on the basis of previous research (+intuition) that can account for the speaker's choice between the cx
- fit a multivariate model to test if these variables help predict the use of cxs

Bottom-up approach

- classify the lexemes that fill in the constructional slots (Cr, Ce, Vinf) on the basis of their distributional properties (Semantic Vector Space models)
- test how well these classifications can predict the choice between *let*, *allow* and *permit*.

PREVIOUS SCHOLARSHIP

Iconicity

- Haiman (1983): the conceptual distance between the cause and the result corresponds to the formal distance between the causative auxiliary and the effected predicate.

Iconicity

- Haiman (1983): the conceptual distance between the cause and the result corresponds to the formal distance between the causative auxiliary and the effected predicate.
- Duffley (1992): *let* + bare Vinf expresses more tightly integrated events than *allow* /*permit* + *to*-Inf.

Iconicity

- Haiman (1983): the conceptual distance between the cause and the result corresponds to the formal distance between the causative auxiliary and the effected predicate.
- Duffley (1992): *let* + bare Vinf expresses more tightly integrated events than *allow /permit* + *to*-Inf.

? Will this factor survive a quantitative test in a multifactorial model, which also contains other semantic, syntactic, collocational, stylistic and social variables? ?

DATA SET AND VARIABLES

Data set

- random samples of *let*, *allow* and *permit* + (…) + (*to*) Vinf from the BNC XML edition

Data set

- random samples of *let*, *allow* and *permit* + (…) + (*to*) Vinf from the BNC XML edition
- excluded:

  - spurious hits

  - passives

  - optatives, hortatives (*let’s go!*) and other non-permissive uses of *let*

Data set

- random samples of *let*, *allow* and *permit* + (...) + (*to*) Vinf from the BNC XML edition
- excluded:

  - spurious hits

  - passives

  - optatives, hortatives (*let's go!*) and other non-permissive uses of *let*

---

882 exemplars × 3 constructions = 2646 observations,

coded for 15 contextual variables

Conceptual integration

- *CeControl*: “X let/allowed/permitted Y (to) do Z, and Y did Z because (s)he chose to do so. “

  - “Yes”: *The professor allowed the students to use their course notes.*

  - “No”: *Let the baby sleep.*

Conceptual integration

- *CeControl*: "X let/allowed/permitted Y (to) do Z, and Y did Z because (s)he chose to do so. "

  - "Yes": *The professor allowed the students to use their course notes.*

  - "No": *Let the baby sleep.*

- *EPVal*: valency of the Infinitive (length of the causation chain)

  - Intransitive: *I let him go*.

  - Transitive: *You allowed me to release them*.

  - Passive Vinf: *You've let him be killed*.

Formal linguistic distance

- *Distance*: formal distance (in words) between Aux and (to)-Vinf

  Live and Let Die (Distance = 0)

  … to permit B., who was anxious to co-operate, to disclose them the documents… (Distance = 6)

Cr and Ce semantics

- *CrSem* and *CeSem:* semantic class of the Causer/Causee
  - *Anim:* animate nouns and pronouns
  - *Mat:* material objects
  - *Abstr:* abstract entities

Channel and domain

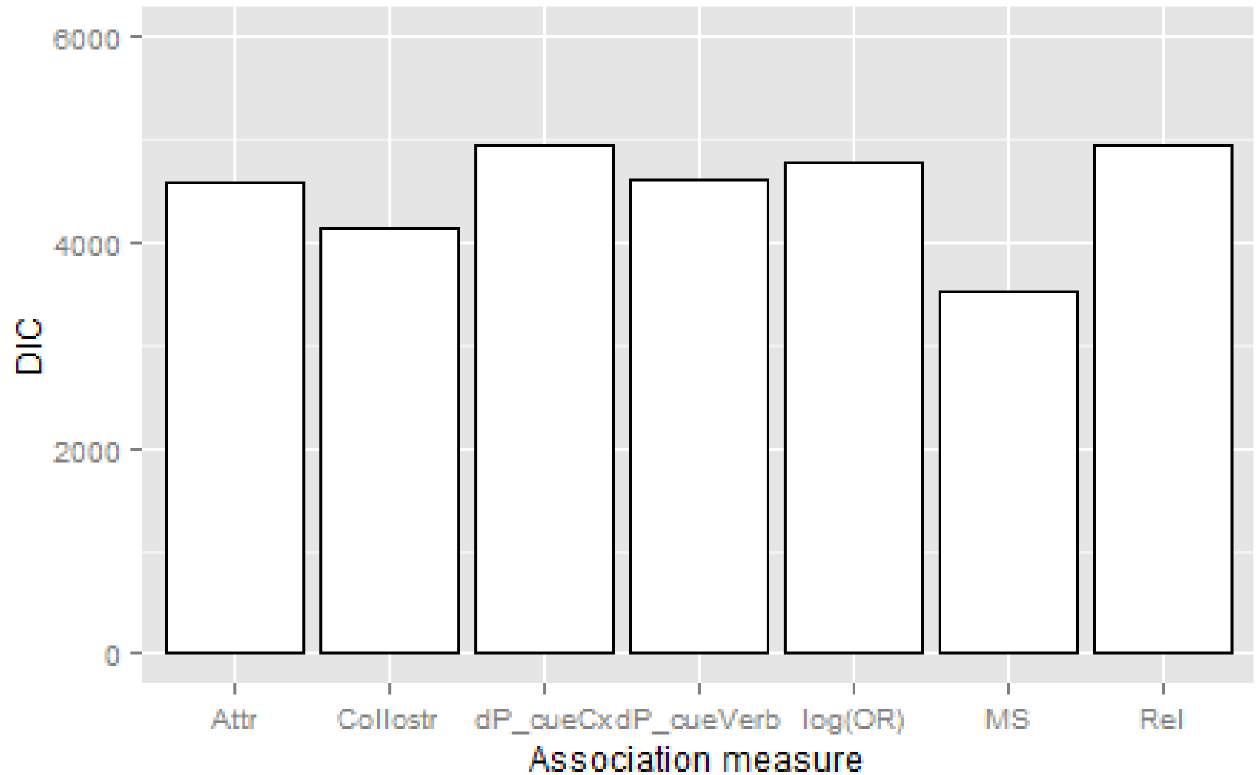
- *Channel*: channel of communication
  - spoken
  - written
- *Domain*: text domain
  - Imagery (fiction)
  - Educational/informative
  - Public (social, political, economic, institutional)
  - Other

Collocational fixation

- Measures of association between each Cx and Vinf:
  - Attraction (Schmid 2010)
  - Reliance (Schmid 2010)
  - Minimum sensitivity (cf. Wiechmann 2008)
  - ΔP (Ellis 2006)
  - log odds ratio
  - Collostructional strength (Stefanowitsch & Gries 2003, etc.)

Represent different relationships (symmetric and asymmetric, with and without contingency information, effect size and significance)

DIC criterion (MCMC glmm)



Minimum sensitivity

$$MS = \min(x, y)\text{, where}$$

- x is the frequency of the infinitive in the construction divided by the total frequency of the construction (i.e. Attraction)
- y is the frequency of the infinitive in the construction divided by the total frequency of the verb (i.e. Reliance)

Stylistic factor: horror aequi

- *Horror_aequi*: avoidance of repetition (cf. Brugman 1909; Rohdenburg 2003). Coded as "Yes" is there is another verb of letting in the left context in the same sentence.

Yet, ITV regards its competition so seriously that it refuses even to allow Sky the traditional news access to its exclusive sponsored events or permit the EBU to pass on its pictures from events such as athletics' European Cup.

Other factors

- *Morph:* TAM characteristics of the letting Auxiliary
- *Coref*: coreferentiality between Cr and other participants

  I have no patience with women who let themselves go.
- *Possess:* possessive relationships between Cr and other participants (marked by a possessive marker)

  Let your letter express your personality.
- *Polarity*:

  - Pos: positive

  - Neg: negative, e.g. *I do not allow a woman to make a fool of me.*
- *Nchar:* length of the Vinf (in characters)
- *Vfreq*: frequency of the infinitive

FULL MCMC GLMM

The model

- Multinomial mixed model
- the response: the letting Cx (let, allow or permit)
- 15 fixed effects (the contextual variables)
- Vinf as random effects
- Bayesian approach, Markov Chain Monte Carlo method, R package MCMCglmm
- 40,000 iterations
- Accuracy: 0.704 (against the baseline of 0.33)

Fixed effects (a selection)

Variable	Coefficient (post. mean)	pMCMC
Distance.allow	0.246	0.004
Distance.permit	0.533	< 0.001
CeControl=Yes.allow	1.18	< 0.001
CeControl=Yes.permit	1.396	< 0.001
MS.allow	-863.4	< 0.001
MS.permit	-4,679	< 0.001
Channel=Written.allow	1.336	< 0.001
Channel=Written.permit	4.03	< 0.001
Nchar.allow	0.371	< 0.001
Nchar.permit	0.475	0.001
CrSem=Abstr.allow	3.17	< 0.001
CrSem=Abstr.permit	3.798	< 0.001

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

‘Harmonic alignment’

LET - ALLOW (/) PERMIT

formal linguistic distance

conceptual distance

social distance, formality

spatiotemporal distance between interlocutors

lack of collocational fixation

less cognitively salient Cr and Ce

BOTTOM-UP APPROACH: SVS CLASSES

Semantic Vector Spaces

- A popular method in distributional approaches to semantics: Lexemes or word forms are represented as vectors of weighted co-occurrence frequencies with contextual features.

Semantic Vector Spaces

- A popular method in distributional approaches to semantics: Lexemes or word forms are represented as vectors with weighted co-occurrence frequencies of contextual features.
- First implementation for CxGr studies on Dutch causative constructions in Levshina & Heylen (Forthc.).

  - We can use different SVS models of constructional collexemes (Cr, Ce and Vinf) to cluster the collexemes and see which classification helps us best predict the use of near-synonymous constructions

Models for Cr and Ce

- Step 1. Create two lists of nouns that are used in the Cr and Ce slots in the entire data set.

Models for Cr and Ce

- Step 1. Create two lists of nouns that are used in the Cr and Ce slots in the entire data set.
- Step 2. Obtain co-occurrence frequencies of the nouns with 5000 most frequent lexemes in the BNC

  - BOW4: all words in the window 4 words on the left from the target noun and 4 on the right.

  - BOW15: the same, but the window is 15.

Models for Cr and Ce

- Step 1. Create two lists of nouns that are used in the Cr and Ce slots in the entire data set.
- Step 2. Obtain co-occurrence frequencies of the nouns with 5000 most frequent lexemes in the BNC

    - BOW4: all words in the window 4 words on the left from the target noun and 4 on the right.

    - BOW15: the same, but the window is 15.

- Step 3. Transform the co-occurrence frequencies into Positive Pointwise Mutual Information scores.

Models for Cr and Ce

- Step 1. Create two lists of nouns that are used in the Cr and Ce slots in the entire data set.
- Step 2. Obtain co-occurrence frequencies of the nouns with 5000 most frequent lexemes in the BNC

  - BOW4: all words in the window 4 words on the left from the target noun and 4 on the right.

  - BOW15: the same, but the window is 15.
- Step 3. Transform the co-occurrence frequencies into Positive Pointwise Mutual Information scores.
- Step 4. Compute the distances between all pairs of word vectors as $dist_{AB} = 1 - \cos(A, B)$.

Models for Cr and Ce

- Step 1. Create two lists of nouns that are used in the Cr and Ce slots in the entire data set.
- Step 2. Obtain co-occurrence frequencies of the nouns with 5000 most frequent lexemes in the BNC

  - BOW4: all words in the window 4 words on the left from the target noun and 4 on the right.

  - BOW15: the same, but the window is 15.
- Step 3. Transform the co-occurrence frequencies into Positive Pointwise Mutual Information scores.
- Step 4. Compute the distances between all pairs of word vectors as $dist_{AB} = 1 - \cos(A, B)$.

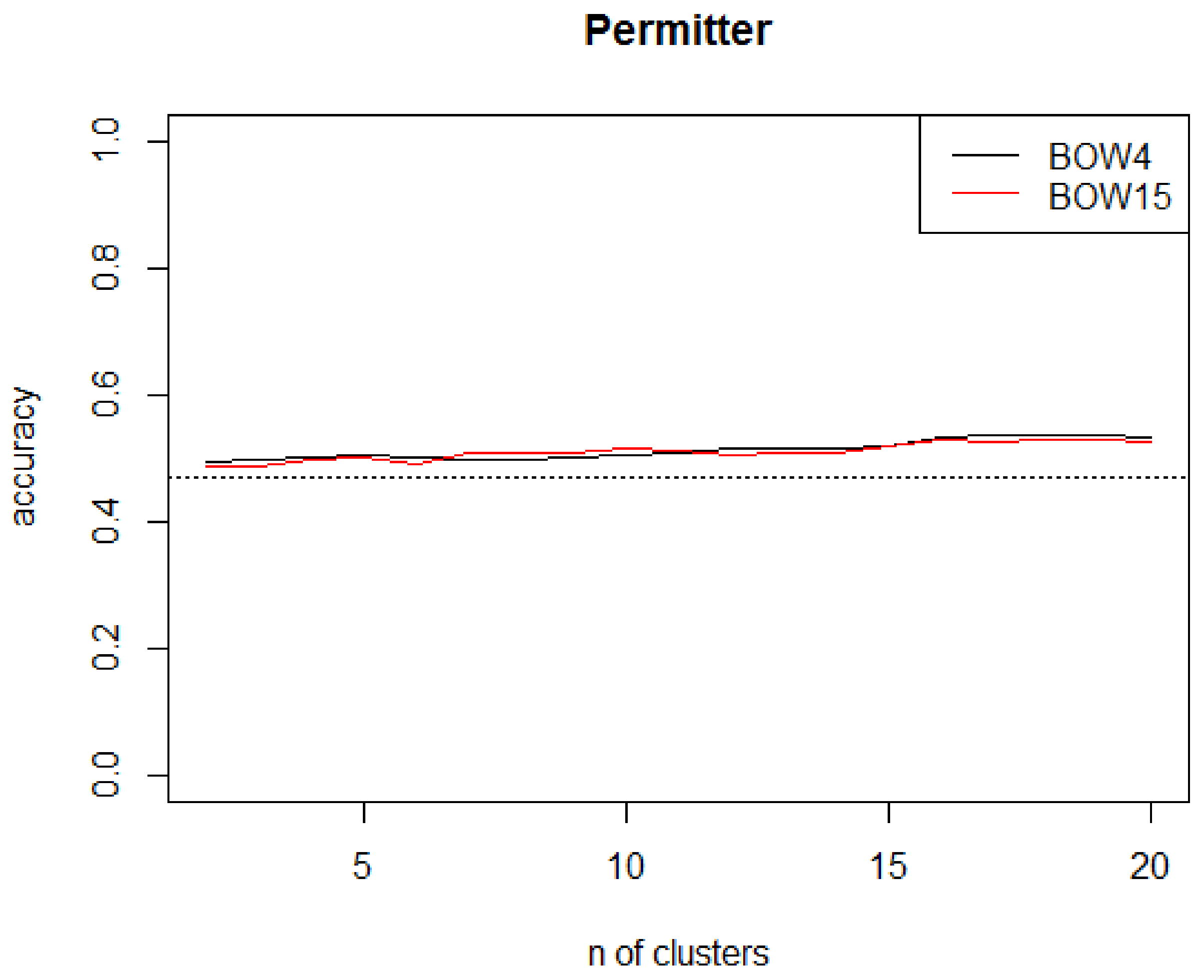
Models for Cr and Ce

- Step 5. Use Partitioning Around Medoids clustering method to classify the nouns into *n* clusters (from 2 to 20).

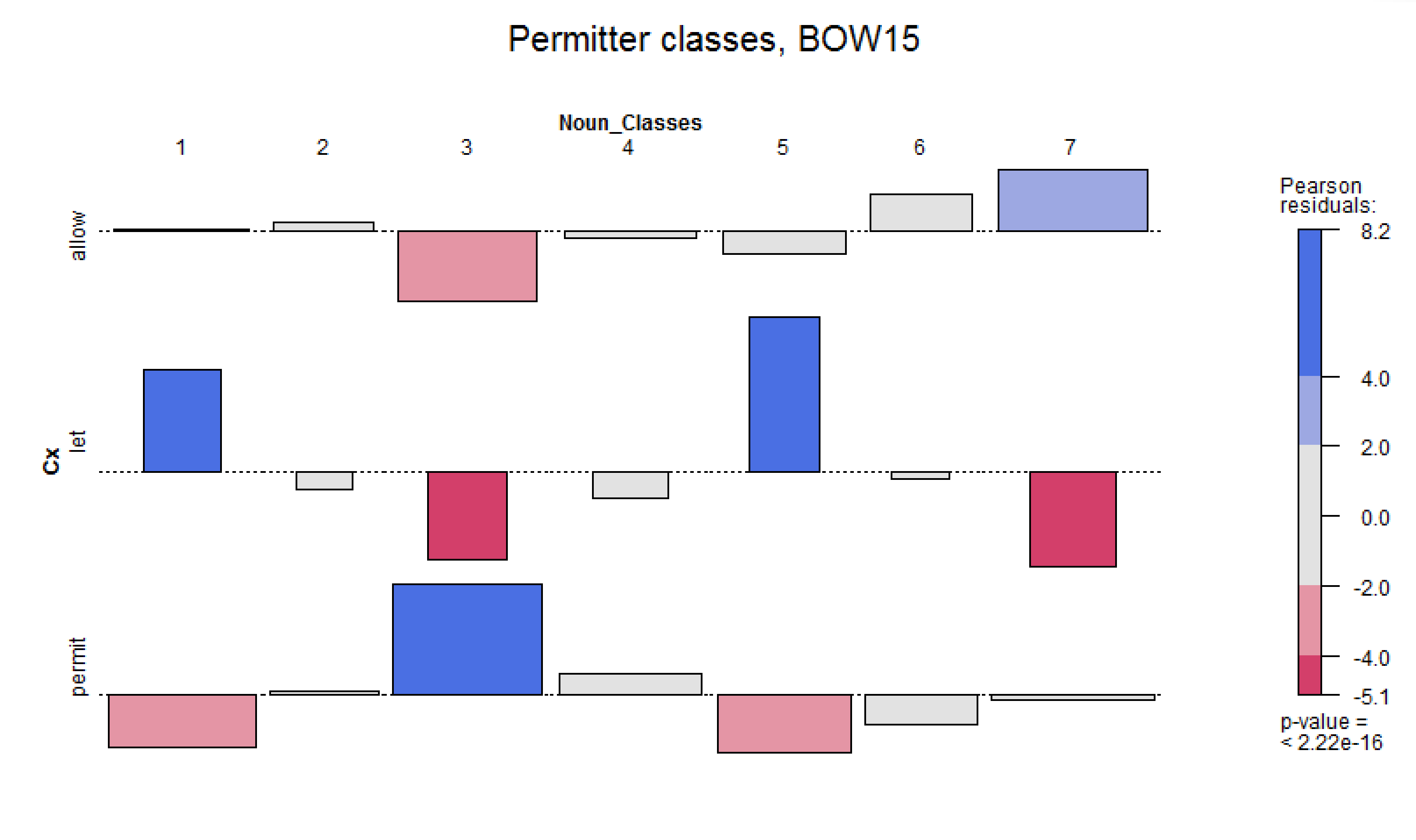
Models for Cr and Ce

- Step 5. Use Partitioning Around Medoids clustering method to classify the nouns into *n* clusters (from 2 to 20).
- Step 6. Compute goodness-of-fit measures of a classification model (SVM, multinomial regression) for every clustering solution, i.e. how well every solution helps us predict the choice between let, allow and permit.

Classes of Cr



Cr: BOW15, 7 clusters



Cr: Interpretation of clusters

- Cluster 1 (pro-*let*)

person's names (John, Juliet, Miranda, Shakespeare, BBC, Bundesbank) and animate common nouns: athlete, boss, countess, student, Masai, successor, champion, king, governor…

- Cluster 3 (pro-*permit*)

legalese: clause, act, argument, certificate, convention, court, document, judge, landlord, law, magistrate, verdict, warrant…

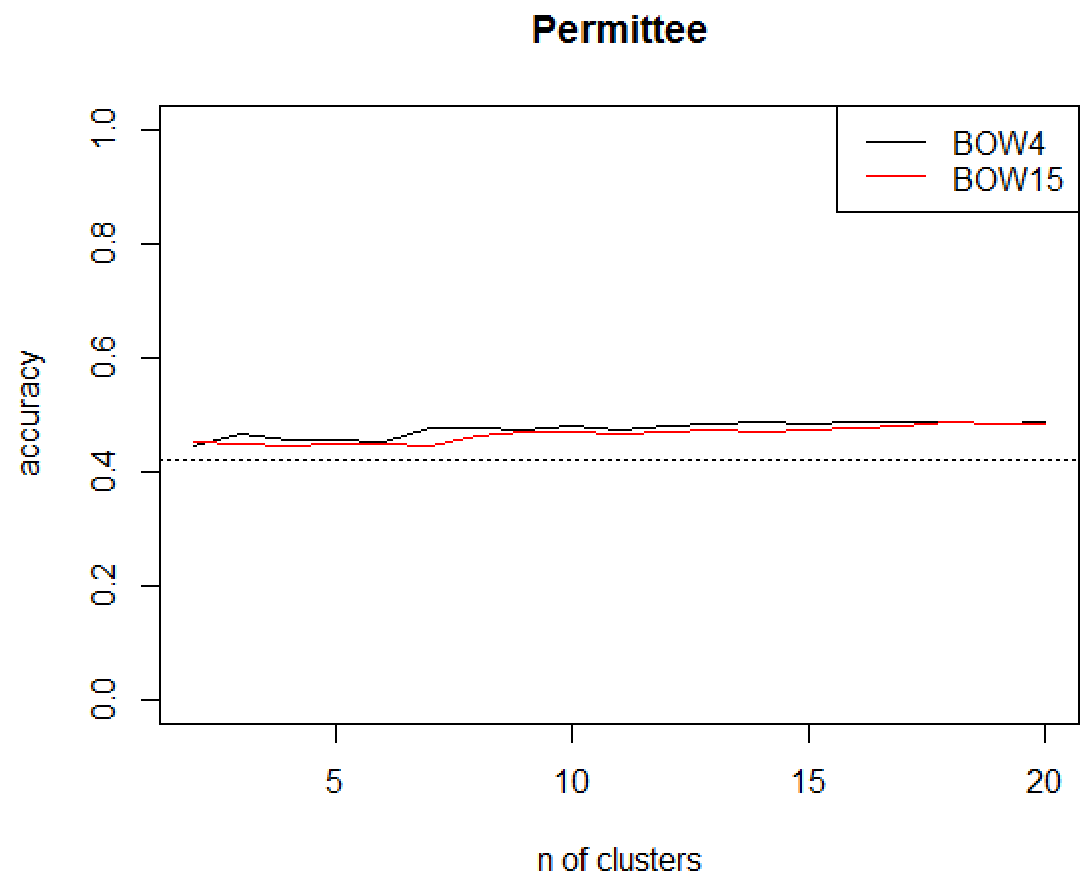
- Cluster 5 (pro-*let*)

animate nouns (esp. family): mother, father, daughter, doctor, female, lad, lady, people, person, pervert, teacher, thief, husband…

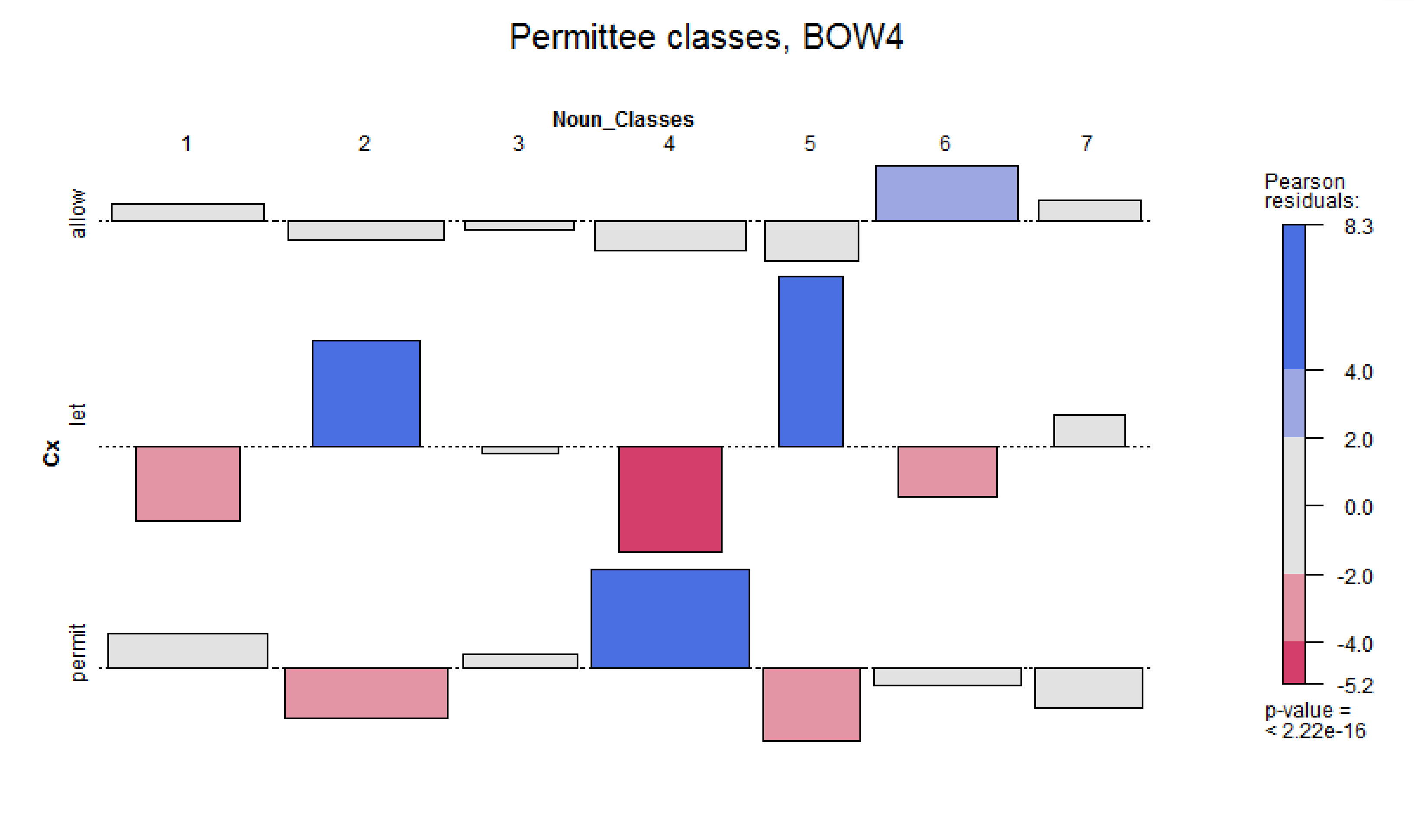
- Cluster 7 (pro-*allow*)

abstract and scientific: addition, adjustment, procedure, result, solution, method, machinery, sampling, science, theory, variation, language, data…

Classes or Ce



Ce: BOW4, 7 clusters



Ce: Interpretation of clusters

- Cluster 2 (pro-*let*)

Humans: American, attacker, daughter, Dane, maniac, murderer, pilot, prince, stranger, youngster…

- Cluster 4 (pro-*permit*)

Public sphere: court, infringement, radical, regime, inspector, resident, republic, suitor, tax, employer, elite, minority, education, department…

- Cluster 5 (pro-*let*)

Physical objects, esp. body parts: arm, bone, breast, eye, finger, foot, hand, head, throat; bed, axe, barbell, blade, pedal, pencil, ring, table…

- Cluster 6 (pro-*allow*)

Abstract entities: beauty, behaviour, belief, bequest, boundary, brokenness, capacity, case, celebration, fertility, feeling…

Cr and Ce: conclusions

- two dimensions:

  - Dim 1: animate – material – abstract

  - Dim 2: Subject domain plays a role (‘general’, science, law/public)

Vinf: SVS models

- Since syntactic information has been crucial for verb classes induction (Schulte im Walde 2009), I fit four models based on syntactic information.
- Based on a segment of the BNC (15%) parsed with Stanford Parser.
- Examples of relations: nsub, dobj, iobj, prep, acomp, ccomp
- Models:

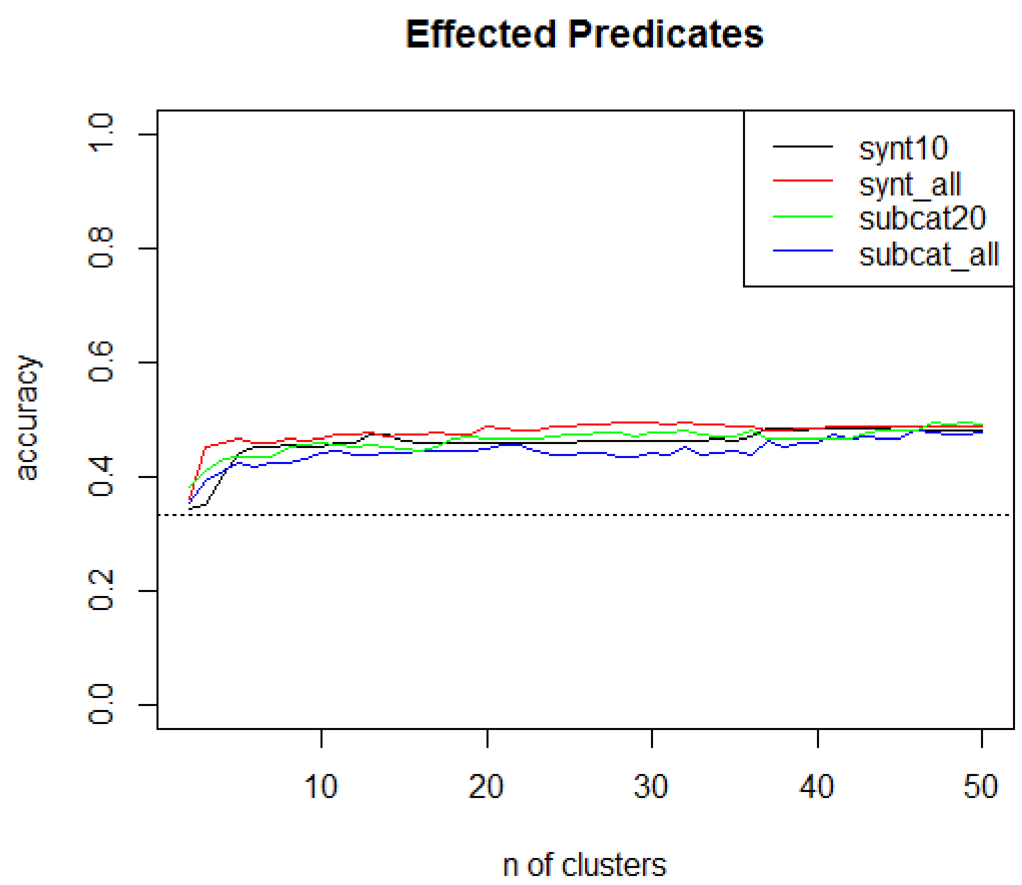
  - synt_all (occurrence in any of syntactic relations where a verb can occur)

  - Synt10 (occurrence in top 10 most frequent relations)

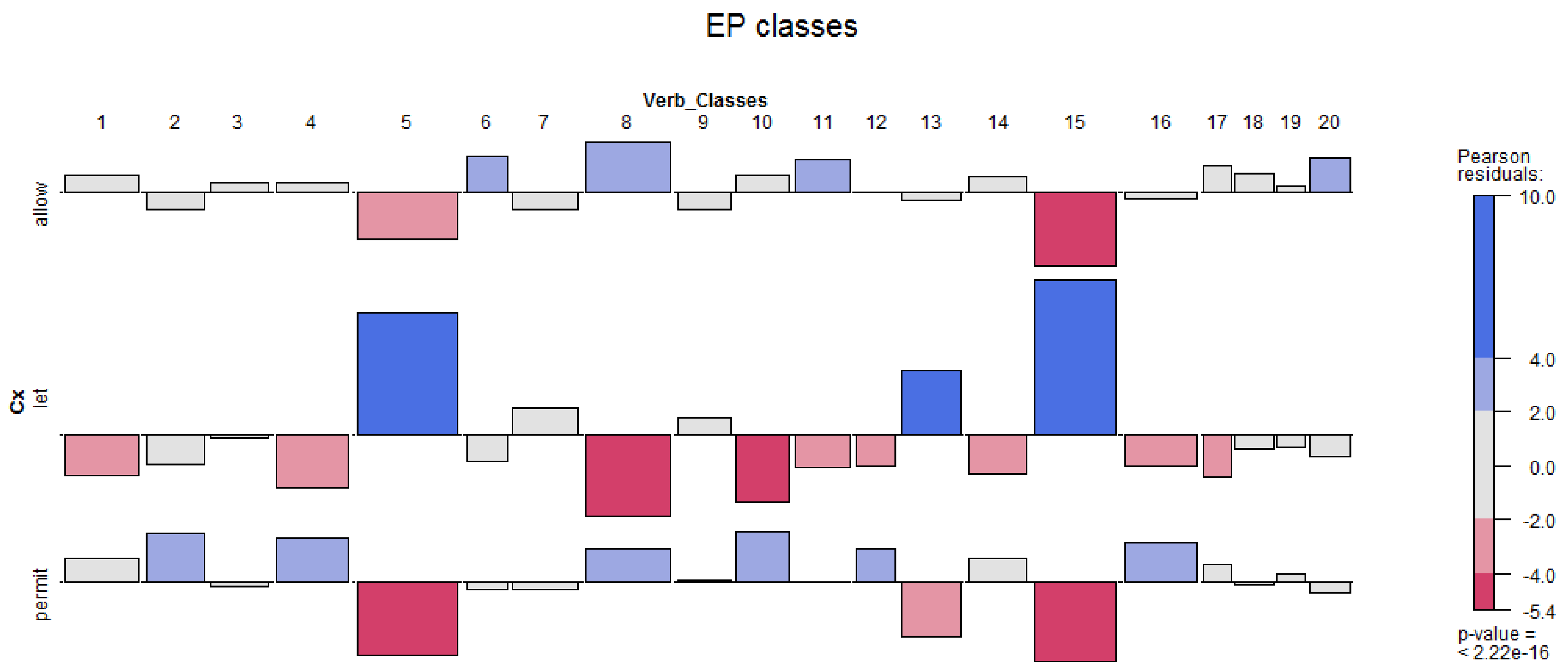
  - subcat_all (occurrence in all possible subcategorisation frames found in the corpus)

  - subcat20 (occurrence in top 20 most frequent subcategorization frames)

Classes of Vinf



Vinf: synt_all, 20 classes



Vinf: Interpretation of clusters

- Cluster 15 (pro-*let*):

perception and other mental verbs: see, know, think, understand, believe, hear, realize, wonder, recall…

- Cluster 5 (pro-*let*):

basic physical states and activities, mostly intransitive: stand, sit, lie; go, walk, come, fly, fall, jump, pass; linger, stay; keep, become…

- Cluster 13 (pro-*let*):

basic causative verbs, physical actions: build, draw, get, break, hold, fold, dig, spill, pick, wash, summon, fasten…

- Cluster 2 (pro-*permit*):

incorporate, relinquish, describe, inspect, reveal, ignore, digress, translate, interpret, suspend, satisfy, delve, engage…

Vinf: conclusions

- Some classes are similar to Levin’s, although not all.
- Again, the domain plays a role.

DISCUSSION

Conclusions: theory

- The iconicity-based theory about the conceptual difference between *let* and *allow/permit* (conceptual cohesion) is supported by the multivariate analysis.

Conclusions: theory

- The iconicity-based theory about the conceptual difference between *let* and *allow/permit* (conceptual cohesion) is supported by the multivariate analysis.
- However, this is only one of many factors that constrain the use of the constructions.

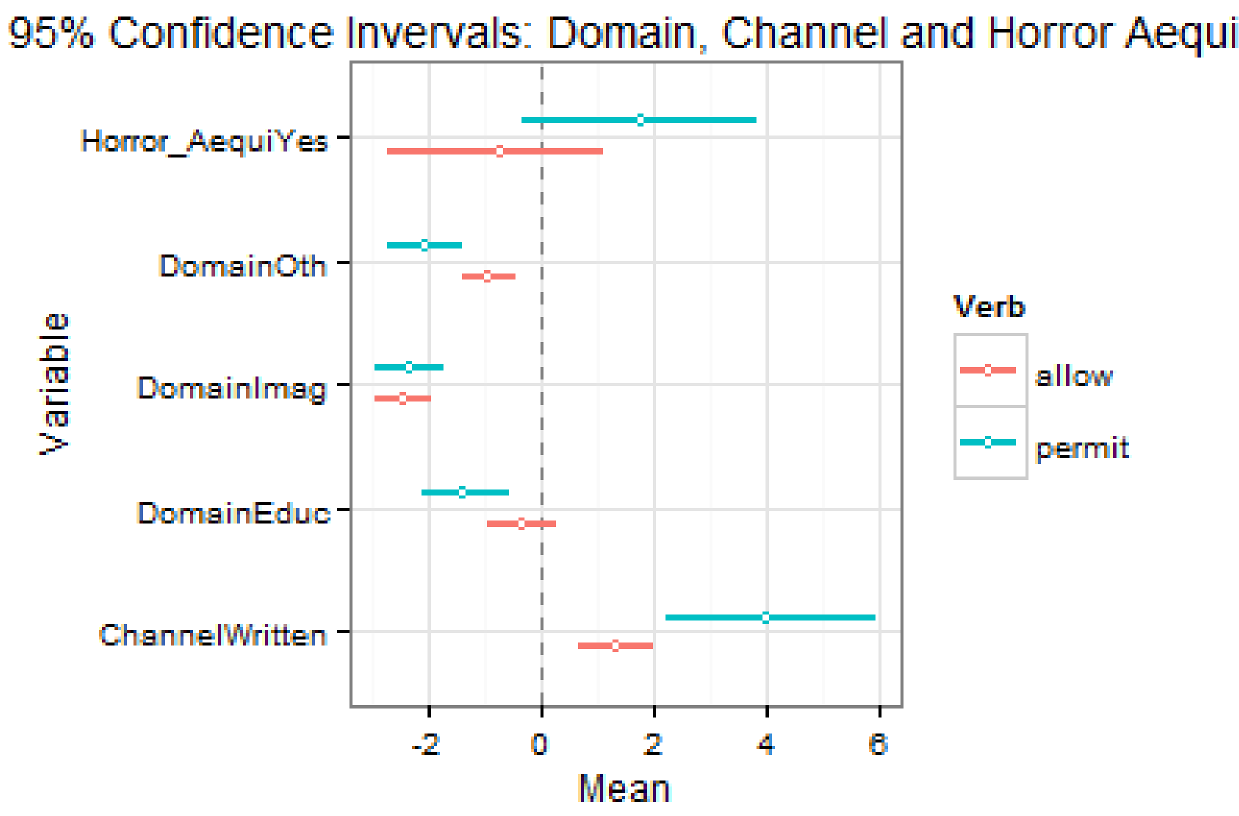
Conclusions: method

- The multidimensionality of the differences between the Cxs is reflected in the distributional SVS-based classifications of constructional collexemes.

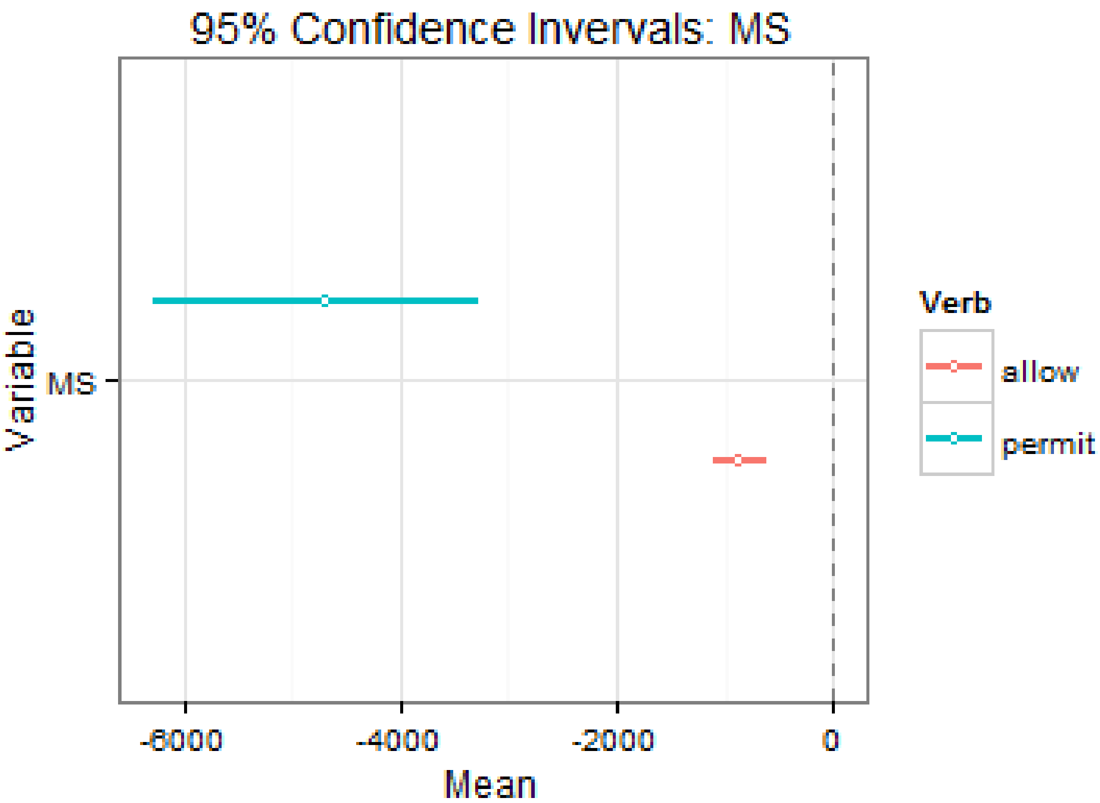
Conclusions: method

- The multidimensionality of the differences between the Cxs is reflected in the distributional SVS-based classifications of constructional collexemes.
- The ‘flat’ clustering approach tested here is probably not the optimal one. One might want to use dimensionality-reduction methods to pin down the relevant dimensions (similar to LSA) and combine these dimensions in a multivariate analysis.

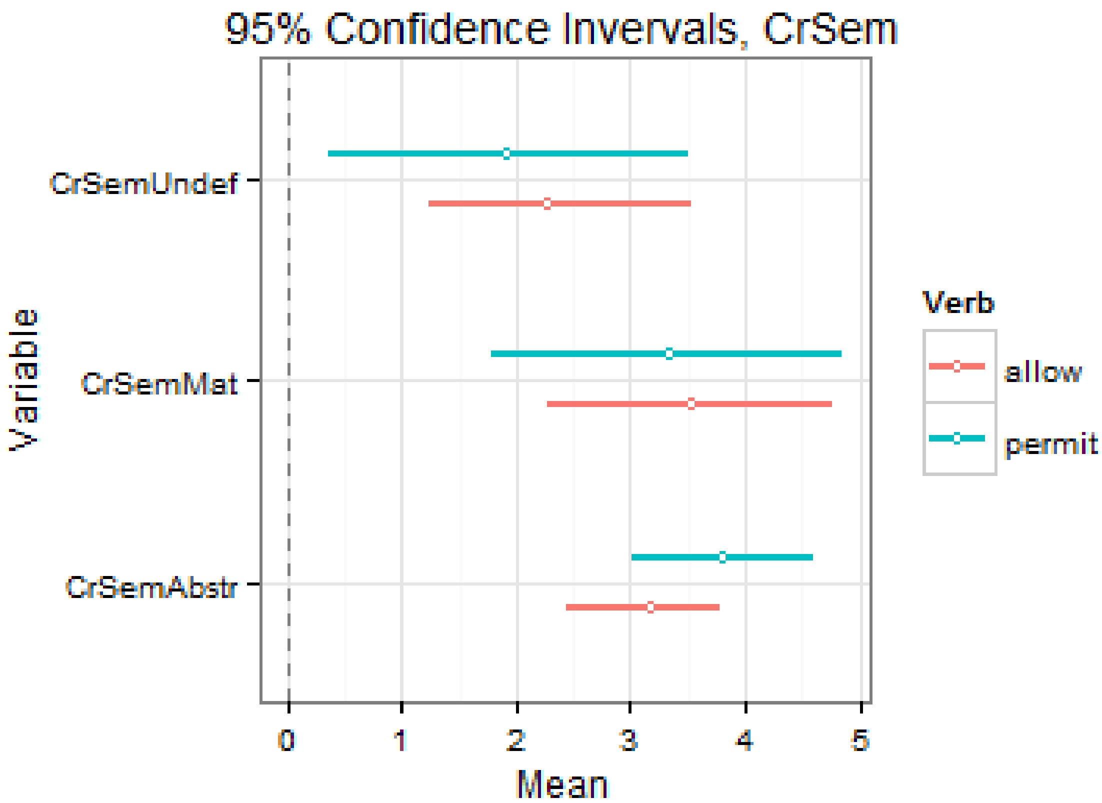
Channel, domain & horror aequi



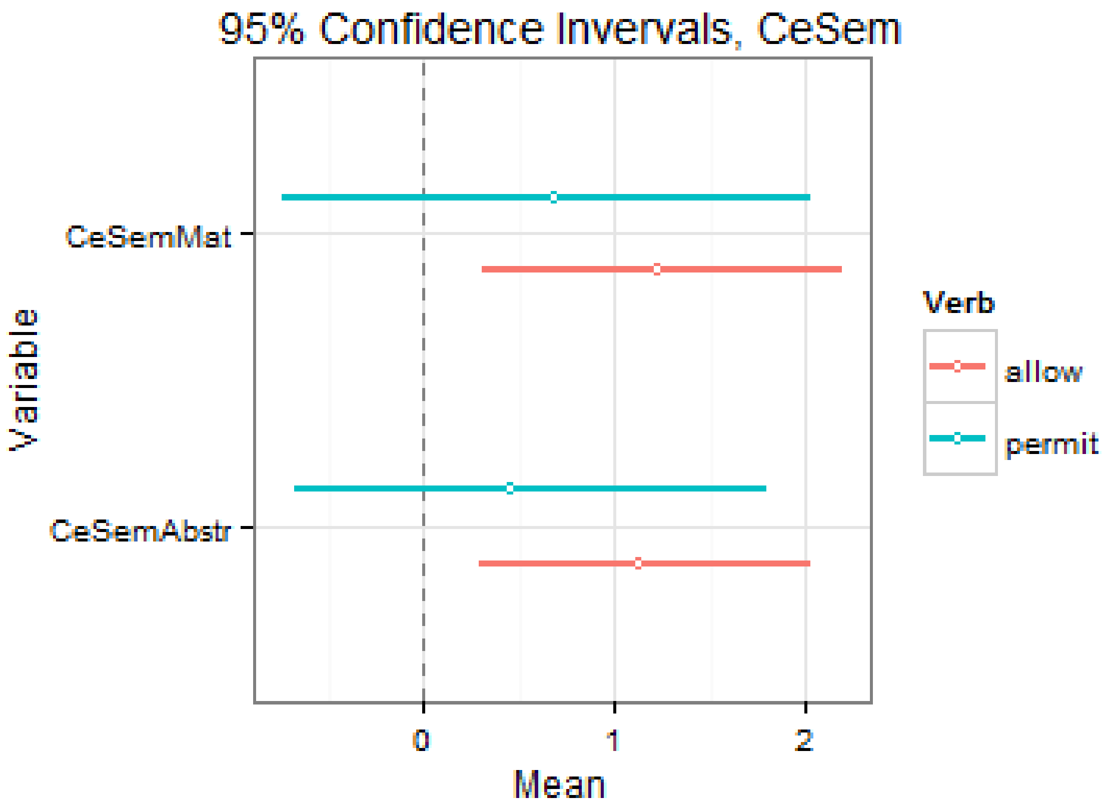
Collocational entrenchment (MS)



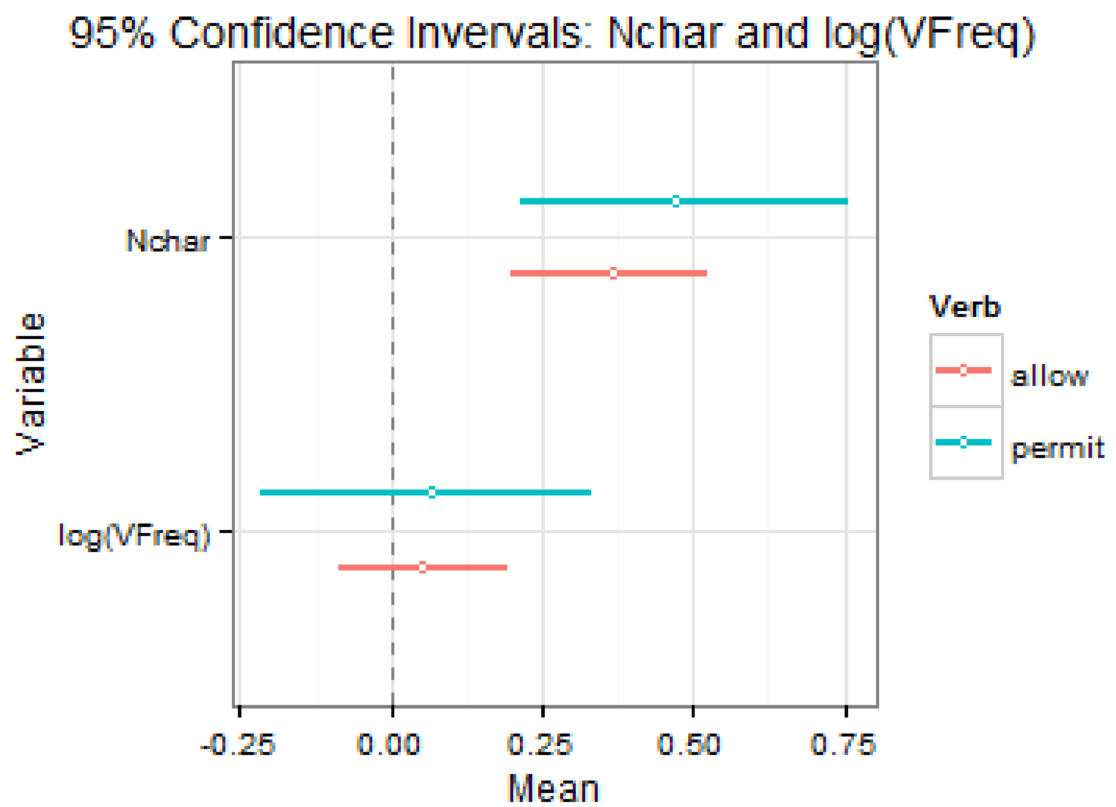
Semantics of Causer



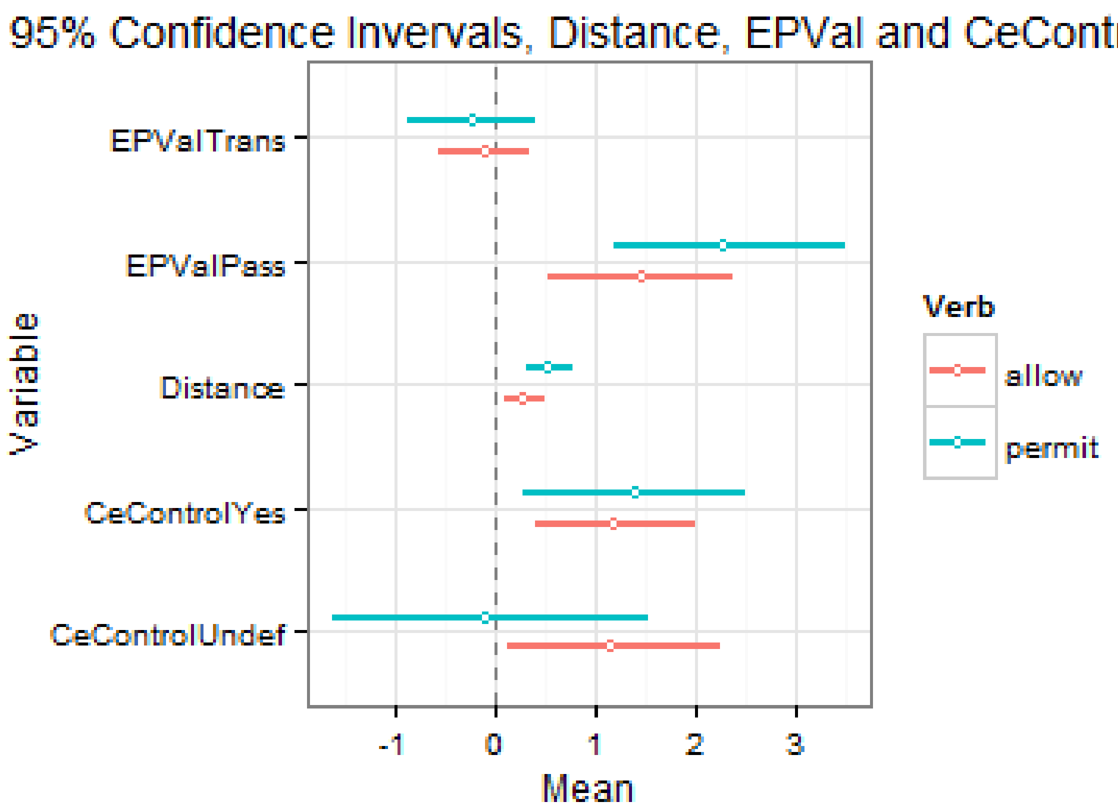
Semantics of Causee



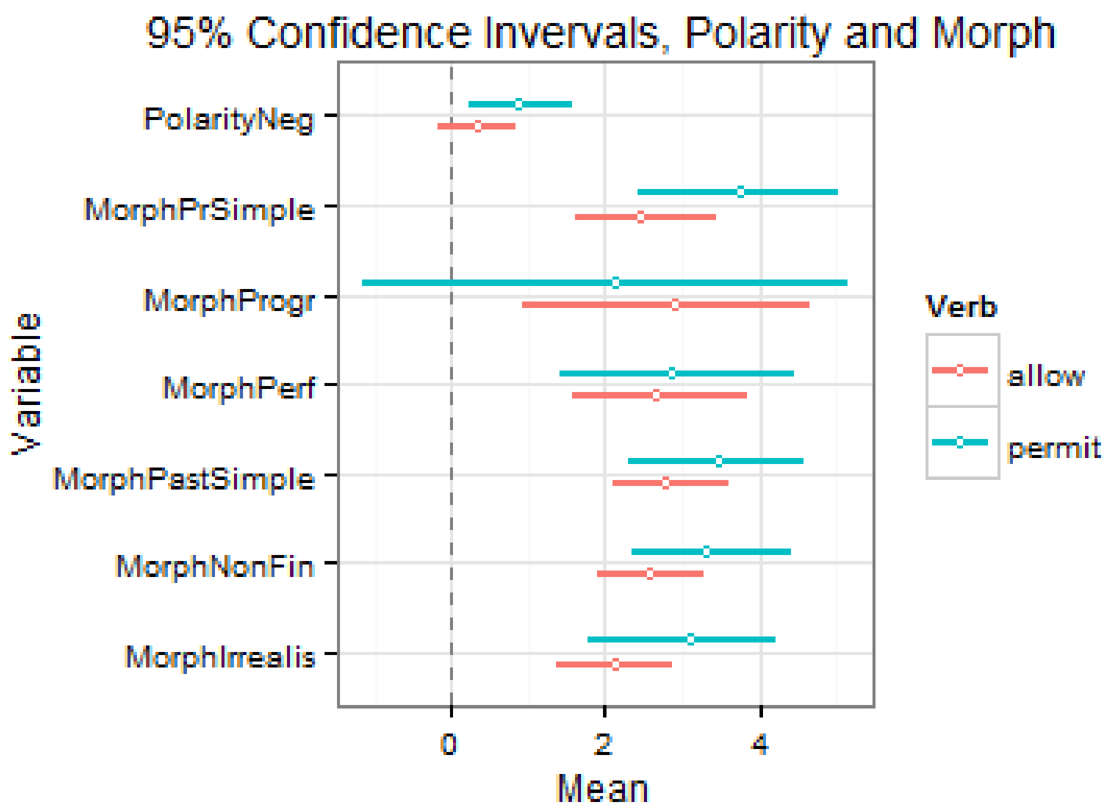
Length and frequency of Vinf



Conceptual Cohesion and Linguistic Distance



Polarity and TAM



MS: Highest-ranking infinitives

Rank	LET	ALLOW	PERMIT
1	go	escape	inspect
2	know	enter	delegate
3	pass	proceed	adduce
4	finish	continue	derogate
5	touch	choose	discontinue
6	forget	develop	quantify
7	happen	operate	re-finance
8	stay	gain	track
9	die	pass	flourish