Annotation of common features for the genitive, dative, and particle placement alternations

Jason Grafmiller, Melanie Röthlisberger,
Benedikt Heller, & Benedikt Szmrecsanyi

Last modified: July 6, 2016

Contents

The following guidelines outline the general annotation process for the genitive, dative, and particle placement alternations in the project "Exploring probabilistic grammar(s) in varieties of English around the world" (EPG).[1] They are based on the coding guidelines developed by Anette Rosenbach (for genitives) and others cited herein.

[1] `http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html`

1 General lay-out of dataframe

At a minimum, the full text forms of the relevant constituents should be included as separate variables (columns) in the dataframe. For the genitive this is the full possessor and possessum, for the dative the full theme and full recipient, and for particle placement the full direct object and the particle. It is also a good idea to include the entire genitive NP or the VP (in the dative and particle placement cases) as its own separate column. This will allow easier coding of alternative predictors, checking for errors, and checking for the values of vaguely named variables. In addition, there should be separate columns containing the head noun (lemma) of each constituent. Each token must also be indexed for the text it occurs in (as a text identifier), the line number within the text, speaker identity (where applicable), region/variety, register (this may require multiple columns), and of course, the outcome variable of construction type produced.

To allow further post-editing for certain features it is desirable to also include a column containing a substantial amount of preceding context (call it **PriorContext**, or something like that). The size of this context will depend on the nature of the annotation. Some common methods use either the 10 lines or 100 words preceding the token in question. Since lines can differ substantially in the nuber of words used, especially across different registers, we will go with the latter method.

Naming conventions

For individuals/teams working on separate dataframes, it is also a good idea to establish some consistent naming conventions to make cross-comparisons easier. Generally, names of columns should be transparent for anyone with a basic familiarity with the data, phenomena, and features under investigation. Since the primary analysis will be done in R, names should not begin with numerals, nor (ideally) should they contain punctuation marks or spaces, as these can create clunky names when imported into R. For example, simple names such as **PorAnimacy**, **PumAnimacy**, **RecDefiniteness**, **ThemeDefiniteness**, or **DirObjGivenness** are nice because they are reasonably short, but still interpretable.

1.1 Columns for metadata

The following are suggested names for the metadata columns. These column names should be the same across all datasets in the PGEWW project. The various ID columns are designed to provide unique identifiers at various grouping levels, e.g. textfile, text within textfile (there can be more than one), sentence within text, speaker within text, etc. Having these separate IDs will be necessary for distinguishing the different (nested) random effects structures that we may explore.

1. **Variety**: English variety of the token (‘CAN’, ‘GB’, ‘HK’, ‘IND’, ‘IRE’, ‘JAM’, ‘NZ’, ‘PHI’, ‘SNG’)

2. **TokenID**: Unique identifier for a token in the dataset. Construction tag (‘GEN’, ‘DAT’, ‘PRT’), followed by the number of the token in the dataset. This can be as attaching a number

from 1 to whatever. All that is needed is a way to uniquely identify specific tokens. (The other IDs do not suffice as there can be mutliple observations even within the same sentence/line).

3. **FileID**: The file in which the token is found. Variety tag with filename, separated by colon ('GB:S1A-005', file S1A-005.txt in the ICE-GB corpus)

4. **TextID**: The number of the text in the file in which the token is found. Variety tag with filename followed by text number, separated by colons ('GB:S1A-005:1' indicates text 1 in file S1A-005.txt of the ICE-GB corpus)

5. **LineID**: The number of the sentence/line in the file in which the token is found. *Note: Line numbers are not grouped by texts within files.* Variety tag with filename followed by line number, separated by colons ('GB:S1A-005:52' indicates line 52 in file S1A-005.txt of the ICE-GB corpus)

6. **SpeakerID**: The identifier of the speaker within a text. Since written texts do not have speakers, authors of individual texts are all coded as 'A'. Variety tag with filename, followed by text number, followed by speaker/author tag, separated by colons ('GB:S1A-005:1:B' indicates speaker B in text 1 in file S1A-005.txt of the ICE-GB corpus)

7. **GenreFine**: Fine-grained 12-level distinction in register corresponding to the file types shown in Table 1.

8. **GenreCoarse**: Four-level register distinction corresponding to the two major register groups for each modality ('dialogue', 'monologue', 'printed', 'non-printed')

9. **Mode**: The spoken/written modality of the token ('spoken', 'written')

10. **PriorContext**: The 100 words preceding the observation, in plain text (we may also want to include tagged version)

11. **PriorContextTag**: The POS tagged version of the prior context

'PrivateDia' (S1A)	'PublicDia' (S1B)	'UnscriptMono' (S2A)
'ScriptedMono' (S2B)	'StudentWrit' (W1A)	'Letters' (W1B)
'AcademicWrit' (W2A)	'PopularWrit' (W2B)	'Reportage' (W2C)
'InstructWrit' (W2D)	'PersuasiveWrit' (W2E)	'CreativeWrit' (W2F)

Table 1: Fine-grained genres in ICE corpora.

1.2 Columns for data

The following are suggested names for the columns containing the data relating to the observations to be annotated and analyzed.

1. **SentencePlain**: The full sentence/line containing the token, in plain text

2. **SentenceTag**: The POS tagged version of the sentence/line

3. **WholeConstruction**: The full construction being analyzed. This is the full genitive NP (e.g. *Mary's favorite chainsaw, the spells of an evil wizard*), and the full VPs for the dative and particle placement tokens (e.g. *give the gorilla a kiss, beamed the Klingons up*)

4. **WholeConstructionTag**: The POS tagged version of the WholeConstruction column

5. **Full content of the constituents**. The exact content of these columns will depend partly on what we decide to do about discourse markers, disfluencies, and possibly other adverbials (see section 2.2). It is possible that there will be multiple columns for each of these.

   (a) Genitives

   - **PossessorPlain**: Full plain text content of possessor NP, e.g. "Mary", "an evil wizard"
   - **PossessumPlain**: Full plain text content of possessum NP, e.g. "favorite chainsaw", "the spells"

   (b) Datives

   - **RecipientPlain**: Full plain text content of recipient NP, e.g. "the gorilla"
   - **ThemePlain**: Full plain text content of theme NP, e.g. "a kiss"

   (c) Particle placement

   - **DirObjPlain**: Full plain text content direct object NP, e.g. "the Klingons"

6. **Tagged full content of the constituents**. The POS tagged version of the constituent columns

   (a) Genitives

   - **PossessorTag**: `Mary_NP1`, `an_AT evil_JJ wizard_NN1`
   - **PossessumTag**: `favorite_JJ chainsaw_NN1`, `the_AT spells_NN2`

   (b) Datives

   - **RecipientTag**: `the_AT gorilla_NN1`
   - **ThemeTag**: `a_AT kiss_NN1`

   (c) Particle placement

   - **DirObjTag**: `the_AT Klingons_NP2`

7. **Heads of the constituents**. Except in the case of verbal constituents, e.g. *He gave up drinking*, these should be the lemmas and not the wordforms.[2] That is, for the two NPs *the car* and *many cars*, the head for each would be coded as simply 'car'.

   (a) Genitives

   - **PorHead**: Head noun of possessor NP, e.g. "Mary", "wizard"

[2]Where possible, lemmatization should be used over stemming (see `http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html`).

- **PumHead**: Head noun of possessum NP, e.g. “chainsaw”, “spell”

(b) Datives

- **RecHead**: Head noun of recipient NP, e.g. “gorilla”
- **ThemeHead**: Head noun of theme NP, e.g. “kiss”

(c) Particle placement

- **DirObjHead**: Head of direct object NP, e.g. “Klingon”

8. **Resp**: The response variable, i.e. the construction chosen in the given token (*s*-genitive vs. *of*-genitive, double object vs. prepositional dative, continuous vs. discontinuous verb-particle placement).

2 Predictors and coding levels

The following predictors are coded for each of the relevent constituents among the three constructions: possessor/possessum, recipient/theme, and direct object.

2.1 Animacy

Following Wolk et al. (2013), **Animacy** is coded in five levels, with separate columns indicating the animacy level of each constituent. Additional columns may also be added for a more coarse-grained animacy classification (‘animate’ vs. ‘inanimate’), but the initial 5-way classification should be coded first.

Animacy

Code	Category	Comments	Examples
'a'	Human & animal	Only higher animals (not e.g. 'fish' or 'bugs'); includes spirits, god(s), and other agentive (human-like) supernatural entities	*Shakespeare, engineers, the horse, a sixteen-year-old girl, Mr. Kennedy, God*
'c'	Collective	Organizations or political states/bodies when seen as having collective purpose, agenda or will	*the House of Lords, the church, parliament, another country*
		Group of animate individuals with potential variable anaphoric reference (*it/they*)	*family, multitudes, the public, a convoy, the majority*
'i'	Inanimate	Non-temporal, non-locative inanimates: concrete and abstract, all gerunds, participles, and infinitives	*the table, oxygen, other topics, drinking*
'l'	Locative	Places qua places, not groups of inhabitants/members, including *state/empire*; not referable by *they*	*the sea, the playground, China, the earth*
't'	Temporal	Noun or adverb with time reference	*yesterday, last week, March, 1986, this morning*

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorAnimacy**: Animacy of possessor
    - **PumAnimacy**: Animacy of possessum
2. Datives
    - **RecAnimacy**: Animacy of recipient
    - **ThemeAnimacy**: Animacy of theme
3. Particle placement
    - **DirObjAnimacy**: Animacy of direct object

2.2 Length (weight)

The length or syntactic weight of the relevant constituents are coded separately in individual columns. Two measures are used for our analyses: length in orthographic words, and length in orthographic

letters (graphemes). A few points about these counts are worth noting:

- For grapheme counts, spaces are included, while all punctuation is excluded.
- Hyphenated compounds are counted as 2 words, and hyphens are ignored when counting graphemes. Contractions are counted as 1 word.
- Different texts may use acronyms (*NASA, NATO*) and initialisms (*U.S.S.R., the U. N.*) in different ways, i.e. with or without full stops (periods) and with or without spaces in between characters. We considered all acronyms and initialisms to constitute 1 word only, regardless of spacing. Similarly, the word length of numbers, e.g. *1999, flight 93*, is counted as 1 word.
- Discourse markers (e.g. *of course, I think, like, um*), are counted as part of the relevant constituent in which they occur.
- Different VoEs use different spelling conventions. Some of these are inconsequential for measuring length (*analyse* vs. *analyze*), while others can potentially affect the resulting measurements (*doughnut* vs. *donut*). We did not correct for variation in spelling across varieties.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorWordLth**: Length of possessor in words
    - **PorLetterLth**: Length of possessor in letters
    - **PumWordLth**: Length of possessum in words
    - **PumLetterLth**: Length of possessum in letters
2. Datives
    - **RecWordLth**: Length of recipient in words
    - **RecLetterLth**: Length of recipient in letters
    - **ThemeWordLth**: Length of theme in words
    - **ThemeLetterLth**: Length of theme in letters
3. Particle placement
    - **DirObjWordLth**: Length of direct object in words
    - **DirObjLetterLth**: Length of direct object in letters

2.3 Definiteness

According to Anette Rosenbach, the coding of definiteness she developed for the genitive alternation was intended to capture some of the effect of givenness. While this may be sufficient for the genitive alternation, as personal pronouns and indefinite possessums are categorically excluded, the literature on the other alternations is somewhat mixed when it comes to the interaction of definiteness and givenness. Bresnan et al. (2007) find significant effects of both definiteness and accessibility (givenness) in the dative alternation, but Bresnan & Ford (2010) find only definiteness to be a significant predictor (p177). Bresnan and Hay (2008), on the other hand, looked only at accessibility, and did not include definiteness in their model(s), while Wolk et al. (2013) included only definiteness in their dative model. The differential roles of definitness and givenness in particle placement are even less understood. Considering that the relative effects of definiteness, givenness, and pronominality are still not very well understood, I suggest we keep our definiteness coding as simple as possible, for the time being.

Definiteness of the possessor, recipient/theme, and direct object is coded according to the following scheme:

Definiteness

Code	Category	Comments	Examples
'def'	Definite NP	Proper nouns and any of the NP types listed in section 2.3.1	*his shoe, the polls, myself, all my money, what you don't want*
'indef'	Indefinite NP	Any of the NP types listed in section 2.3.2	*a new language, people, some elderflower cordial*

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorDefiniteness**: Definiteness of possessor
    - **PumDefiniteness**: Definiteness of possessum
2. Datives
    - **RecDefiniteness**: Definiteness of recipient
    - **ThemeDefiniteness**: Definiteness of theme
3. Particle placement
    - **DirObjDefiniteness**: Definiteness of direct object

The following subsections provide explicit guidelines for annotating definiteness.

2.3.1 Definite NPs

The following are all the types of NPs that should be coded as 'def' (see Garretson et al. 2004)

- Proper nouns (see section 2.4.1)
- NP with a definite determiner
  - Articles: *the*
  - Demonstrative: *this, that, these, those*
  - Possessive: *her, his, its, my, our, their, your*
  - Quantifier: *all, both, each, either, every, most, neither*
- Definite Pronoun
  - Personal: All, including reflexives and possessives (*mine, hers*, etc.)
  - Impersonal: *each other, everybody (else), everyone (else), everything (else), one another*
  - Wh-pronouns: *which, who, whatever, whatsoever, whichever, whoever, whosoever, whosever*
- An *s*-genitive NP (*George Clooney's bushy beard*)
- Superlatives (*the sourest beer imaginable*)
- Temporal expressions
  - years (*1993*)
  - dollar amounts (*$179000, $20 million*
  - *today, yesterday, tomorrow*
  - *last* or *next* followed by *night, week, month, year*, or any noun referring to a specific day or period of time (e.g. *Easter, March, winter, term, Sunday*)

2.3.2 Indefinite NPs

The following are all the types of NPs that should be coded as 'indef' (see Garretson et al. 2004).

- NP with an indefinite determiner
  - Articles: *a, an*
  - Quantifier: *another, any, enough, few, fewer, half, less, little, little or no, lots of, many, more, much, no, no more, no such, none one-half, one-third (...), one, one or more, ones, plenty of, several, some, twice*
- Indefinite pronouns
  - *any one (else), anybody (else), anyone (else), anything (else), no-one, no-body (else), nothing (else), one's, oneself, somebody (else), someone (else), something (else)*

- Bare plural NPs
- Numbers that are not years or monetary amounts
- Gerunds NOT headed by definite determiners ("screaming" in *the cause of the baby's screaming* is 'def'; but "drinking" in *gave up drinking* is 'indef')
- Any determinerless noun ending in *-tion, -ment, -sion, -ology,* or *-ism*

2.4 NP expression type

To distinguish pronominality (among other things) from the effects of definiteness and givenness, I have added another column describing the syntactic category of the relevant constituent heads. For the most part, these features can be coded mostly automatically, and having such a column will aid in the use of the data later on.

The follow classes of NP expression types should be coded in for each of the relevant constituents in a separate column. Personal pronouns are marked in their own class separate from impersonal pronouns for several reasons. First, while the former are known to behave almost categorically in the three constructions (esp. the genitive and particle placement alternations), the latter are not as categorically restricted.

(1) a. Cars come back at end of reception to pick everyone up and drive them home.

b. The planes come in and pick up everyone, …

Second, impersonal pronouns can be modified (2), unlike personal pronouns.

(2) a. Do you pick up everyone who hails your cab?

b. This will give anyone in Iowa a lump in their throat.

Finally, impersonal pronouns vary with respect to definiteness, unlike personal pronouns which are always definite. The full coding list for NP types is shown below.

NPType

Code	Category	Comments	Examples
'nc'	Common noun	Common noun	*birds, the market, wisdom, this year*
'np'	Proper noun	See section 2.4.1	*President Kennedy, Japan, the United Nations*
'pprn'	Personal pronoun	Personal pronouns, incl. possessives and reflexives.	*me, theirs, yourself*
'iprn'	Impersonal pronoun	Any definite or indefinite pronoun, incl. *wh* pronouns	*everyone, something, whoever*
'dm'	Demonstrative	Bare demonstrative	*this, that, these, those*
'ng'	Gerund	Present participle *-ing* forms (rare)	*give up drinking, hunting's purpose, Give your writing a break*

2.4.1 Proper nouns

It can sometimes be tricky to decide whether a nominal is proper or not. Here is a working test:

- An NP without a determiner (e.g., *Texas*) is proper if it cannot be changed in number or take a determiner (**Texases, *a Texas*). An NP with a determiner (e.g., *the West Indies, A Separate Peace*) is proper if it cannot be changed in number or lose its determiner (**a West Indy, *go to West Indies*). If number or determiner alternation is possible, it is not functioning as a proper noun, and should be treated as a common noun.
- Nouns that are usually proper can be "coerced" into behaving like common nouns, as in *Do you mean the Washington on the Pacific or the Washington on the Potomac?* or *She wants to be a Shakespeare*. In these sentences, the names *Washington* and *Shakespeare* uncharacteristically occur with a determiner, and we therefore say that in this case, they are being **used** like common nouns, not proper nouns. We code such proper nouns used like common nouns as common nouns, since it is actual instances of usage that we are concerned with.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorType**: Type of possessor
    - **PumType**: Type of possessum

2. Datives
    - **RecType**: Type of recipient
    - **ThemeType**: Type of theme
3. Particle placement
    - **DirObjType**: Type of direct object

2.5 Information status

While there are many possible degrees of discourse accessibility, or 'givenness', that could be explored, only two levels of givenness are coded for all three constructions, based on the work by Bresnan & Hay (2008: 249).

Givenness

Code	Comments
'given'	A constituent is coded as 'given' if its referent is mentioned at any time in the 100 words preceding the token in the discourse, or if it is a 1st or 2nd (or 3rd?) person pronoun
'new'	Any constituent that does not refer to a speech participant, and is not referred to in the preceding 100 words is coded as 'new'

Again, a column containing a substantial amount of preceding context will be necessary for checking coreferentiality.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorGivenness**: Givenness of possessor
    - **PumGivenness**: Givenness of possessum
2. Datives
    - **RecGivenness**: Givenness of recipient
    - **ThemeGivenness**: Givenness of theme
3. Particle placement
    - **DirObjGivenness**: Givenness of direct object

2.6 Persistence/priming

For each token, there is a column associated with the persistence measure (Szmrecsanyi 2005). It is labeled as follows:

- **PrimeType**: Type of construction used in the previous choice context (A or B), or 'none' when there is no preceding construction in the 100 words prior to the target construction.

For spoken dialogues, persistence is coded within and across turns, and within and across speakers. The first construction in each conversation or text is automatically to be coded as 'none'. It is important to consider only preceding tokens found in genuine choice contexts, ignoring any occurrences that have been excluded from the analysis.

2.7 Type/token ratio

The lexical density of the surrounding context of the token is estimated using the type-token ratio of the 100 words surrounding the token.

- **TypeTokenRatio:** The type-token ratio is calculated for the 50 words preceding, and 50 words following each construction (Hinrichs & Szmrecsanyi 2007: 457)

The type-token ratio is defined as the number of unique lemmas divided by the number of word tokens in this 100 word environment surrounding the construction in question.

2.8 Frequency

Since we have little information regarding lexical frequency in outer circle varieties of English, standard lexicons (CMU, CELEX) are not ideal. Lemma frequencies for each variety can be obtained from the GloWbE corpus when it is available. The global frequency of a head word (as opposed the text frequency, see below) is normalized as count per million words in the given variety in the GloWbE corpus. Frequency is calculated for the head noun of each of the relevant constituents, and the head noun frequencies of the respective constituent are coded in separate columns.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorHeadFreq**: Frequency of possessor head
    - **PumHeadFreq**: Frequency of possessum head
2. Datives
    - **RecHeadFreq**: Frequency of recipient head
    - **ThemeHeadFreq**: Frequency of theme head
3. Particle placement
    - **DirObjHeadFreq**: Frequency of direct object head

2.9 Thematicity

Thematicity is measured as the normalized text frequency of the head noun in the relevant constituent, i.e. number of uses of the constituent head word/lemma in a text divided by the total number of words in the text (Hinrichs & Szmrecsanyi 2007: 450-451). The normalized text frequency is calculated for the head noun of each of the relevant constituents, and the frequencies of the respective constituent are coded in separate columns.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorThematicity**: Thematicity (text frequency) of possessor head
    - **PumThematicity**: Thematicity of possessum head
2. Datives
    - **RecThematicity**: Thematicity of recipient head
    - **ThemeThematicity**: Thematicity of theme head
3. Particle placement
    - **DirObjThematicity**: Thematicity of direct object head

2.10 NP Complexity

Complex NPs are those that involve any kind of complement and/or postmodification. A fine-grained coding scheme for NP complexity is described below, however a courser distinction could be made between ‘simple’ vs. ‘complex’. While a coarser-grained factor may be more appropriate, there is no less work involved in coding such a factor, since all the types below must be identified as ‘complex’ anyway, thus we might as well mark the finer distinctions on a first pass. [That said, I’m open to a simpler coding system if you have one in mind.]

Complexity

Code	Category	Comments	Examples
'co'	Coordinated NP	Noun phrases involving multiple heads joined with *and, or, but, though,* or any other conjunction	*the onions and the potatoes, Accounting or Economics, silt and floodwaters*
'cp'	Sentential complement	Complement clauses of nouns that take sentential complements; can have overt or null relativizer	*rumors that Obama was not born in the U.S.*
'gn'	Genitive	NP with either an *s-* or *of* genitive	*my father's gun, the cause of all the women*
'pp'	Prepositional phrase	Any PP that is unambiguously modifying the constituent NP (and not some larger constituent, e.g., the VP); this includes non-genitive *of* -PPs	*the lies about Obama, research on these writers, that line of work, his example of the Temperance Society*
'pt'	Non-finite participle	Non-finite VP headed by a participle. Note these are **not** the same as NPs headed by gerunds.	*rule out restricting soviet jewish emigration, hold off pursuing a professional career*
'rc'	Finite relative clause	Finite, restrictive relative clauses. These can have overt or null relative pronouns.	*the guy that caused the accident, the toys you thought were our favorites*
's'	Simple	Any pronoun or NP with [(Det) (A) N] structure. This includes NPs headed by gerunds.	*subscriptions, today, any old rubbish, her head, its previous accomplishments, it, anyone else, his singing*
'vp'	Reduced relative clause	Reduced relatives headed by either present or past participles	*the one sitting on the log, the package damaged by the carrier, the point you made about a possible glut of graduates*

Using the POS-tagged text for each constituent, it is fairly simple to automatically identify complex NPs by the presence of a complementizer, relativizer, verb, preposition, or a genitive *'s*.

The columns for the three datasets are labeled as follows:

1. Genitives
    - **PorComplexity**: Complexity of possessor
    - **PumComplexity**: Complexity of possessum
2. Datives

- **RecComplexity**: Complexity of recipient
- **ThemeComplexity**: Complexity of theme

3. Particle placement
    - **DirObjComplexity**: Complexity of direct object

3 Notes on constituent structure

Here we note some of the decisions that were made about how the relevant constituents were identified, and how the dataframe was constructed. See the construction-specific guidelines for more details.

- **General extenders:** General extenders are included in the dataset in the 'long' column (e.g. PorLong) and excluded in the 'short' column (e.g. PorShort).
- **Incomplete constituents**: Incomplete constituents lack modifying elements. These are usually indicated in the data as <one word> or <several words> and only included if they unambiguously form part of the object constituent. We still need to resolve how to count those words in terms of length.
- **Self-corrections** & **repetitions**: We opted to always go for the first completed overt construction. In case of self-corrections, we use and annotate the corrected version (usually overtly different from the first false start). Repetitions are always included within the constituent in the respective 'long' column but excluded in the 'short' column.
- **Discourse markers:** In cases where intervening material occurs at a constituent boundary (*I gave [the book], um, [to Sam]*), it is counted as part of the following constituent. Separate columns for each weight measure are added to exclude the extra material. This will allow users of the dataframe to decide later how to use the length measures in their analysis.
- **Attachment ambiguities:** Words or phrases that can modify either the NP or the VP (*I picked up the chair in the living room*) are not considered part of the constituent.

References

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. doi:10.1353/lan.0.0189.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118(2). 245–259. doi:10.1016/j.lingua.2007.02.007.

Garretson, Gregory, M. Catherine O'Connor, Barbora Skarabela & Marjorie Hogan. 2004. Coding practices used in the project Optimality Typology of Determiner Phrases. `corpus.bu.edu/documentation/BUNPCorpus_coding_practices.pdf`.

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3). 437–474. doi:10.1017/S1360674307002341.

Shih, Stephanie, Jason Grafmiller, Richard Futrell & Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. In Ralf Vogel & Ruben Vijver (eds.), *Rhythm in Cognition and Grammar*, 207–234. Berlin, München, Boston: DE GRUYTER.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1. 113–149.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419. doi:10.1075/dia.30.3.04wol.