Case and inflection classes: Testing the Form-Frequency-Hypothesis

Matías Guzmán Naranjo & Laura Becker

08.03.2017, 39. Jahrestagung der DGfS, Saarbrücken

The Form-Frequency-Hypothesis

(1) Whenever a category A is expressed by more phonological material than category B, then category A is less frequent than category B.

Whenever category A is more frequent than category B, then category A will be expressed by a shorter or equally long form than category B.

The FFH as a causal implication between frequency and form requires that:

- **ALL** frequency asymmetries result in coding asymmetries in the expected direction, or at least in no asymmetry.
- **ALL** coding asymmetries also have a frequency asymmetry in the expected direction, or at least no asymmetry.

This in turn implies that

- the form-frequency effect must but observable in individual languages.
- Otherwise, observing form-frequency effects cross-linguistically could be due to a correlation with no direct causal relation (i.e. both phenomena follow from a different explanation).

Why case and inflection classes?

What we want to test:

- Do language-specific coding asymmetries correlate with the expected frequency asymmetry?
- Can this correlation be observed across languages?

NB: We want to avoid the confound of semantic complexity and ‘markedness’.

Therefore, **case** and, more specifically, **inflection classes** are a suitable testing ground:

- There is no straightforward reason why, a given case should be semantically more complex than another one.
- Inflection classes have markers of different lengths withing a single language.
- Inflection classes have different frequencies.
- Cases and their exponents are easy to search for in an annotated corpus.

Inflection classes

- 'An inflectional class is a set of lexemes whose members each select the same set of inflectional realizations.' (Aronoff, 1994, p. 64)
- Inflection classes are morphomic, i.e. they are only relevant to morphology, whereas phonology, syntax, or semantics are not sensitive to them.

An example: Gothic *-a* declension

	Day				Word			
	Singular		Plural		Singular		Plural	
Nom	dags	−s	dagōs	−ōs	waúrd	−	waúrda	−a
Acc	dag	−	dagans	−ans	waúrd	−	waúrda	−a
Gen	dagis	−is	dagē	−ē	waúrdis	−is	waúrdē	−ē
Dat	daga	−a	dagam	−am	waúrda	−a	waúrdam	−am

Gothic inflection classes, *-ja* declension

	army				herdsman			
	Singular		Plural		Singular		Plural	
Nom	harjis	−jis	harjōs	−jōs	haírdeis	−eis	haírdjōs	−jōs
Acc	hari	-i	harjans	−jans	haírdi	-i	haírdjans	−jans
Gen	harjis	−jis	harjē	−jē	haírdeis	−eis	haírdjē	−jē
Dat	harja	−ja	harjam	−jam	haírdja	−ja	haírdjam	−jam

	race			
	Singular		Plural	
Nom	kuni	-i	kunja	−ja
Acc	kuni	-i	kunja	−ja
Gen	kunjis	−jis	kunjē	−jē
Dat	kunja	−ja	kunjam	−jam

Gothic inflection classes, *-ō* declension

	gift			
	Singular		Plural	
Nom	giba	−a	gibōs	−ōs
Acc	giba	−a	gibōs	−ōs
Gen	gibōs	−ōs	gibō	−ō
Dat	gibái	−ái	gibōm	−ōm

Gothic inflection classes, *-jō* declension

	band			
	Singular		Plural	
Nom	bandi	−i	bandjōs	−jōs
Acc	bandja	−ja	bandjōs	−jōs
Gen	bandjōs	−jōs	bandjō	−jō
Dat	bandjái	−jái	bandjōm	−jōm

Gothic inflection classes, *-i* declension

	stranger				wife			
	Singular		Plural		Singular		Plural	
Nom	gasts	−s	gasteis	−eis	qēns	−s	qēneis	−eis
Acc	gast	−	gastins	−ins	qēn	−	qēnins	−ins
Gen	gastis	−is	gastē	−ē	qēnáis	−áis	qēnē	−ē
Dat	gasta	−a	gastim	−im	qēnái	−ái	qēnim	−im

Gothic inflection classes, -*u* declension

	son				property	
	Singular		Plural		Singular	
Nom	sunus	−us	sunjus	−jus	faíhu	−u
Acc	sunu	−u	sununs	−uns	faíhu	−u
Gen	sunáus	−áus	suniwē	−iwē	faíháus	−áus
Dat	sunáu	−áu	sunum	−um	faíháu	−áu

Gothic inflection classes, *-an* declension

	man				heart			
	Singular		Plural		Singular		Plural	
Nom	guma	−a	gumans	−ans	haírtō	−ō	haírtōna	−ōna
Acc	guman	−an	gumans	−ans	haírtō	−ō	haírtōna	−ōna
Gen	gumins	−ins	gumanē	−anē	haírtins	−ins	haírtanē	−anē
Dat	gumin	−in	gumam	−am	haírtin	−in	haírtam	−am

Gothic inflection classes, *-ōn* declension

	tongue			
	Singular		Plural	
Nom	tuggō	−ō	tuggōns	−ōns
Acc	tuggōn	−ōn	tuggōns	−ōns
Gen	tuggōns	−ōns	tuggōnō	−ōnō
Dat	tuggōn	−ōn	tuggōm	−ōm

Gothic inflection classes, *-ein* declension

	wisdom			
	Singular		Plural	
Nom	frōdei	−ei	frōdeins	−eins
Acc	frōdein	−ein	frōdeins	−eins
Gen	frōdeins	−eins	frōdeinō	−einō
Dat	frōdein	−ein	frōdeim	−eim

Gothic inflection classes, *-r* declension

	sister			
	Singular		Plural	
Nom	swistar	−ar	swistrjus	−rjus
Acc	swistar	−ar	swistruns	−runs
Gen	swistrs	−rs	swistrē	−rē
Dat	swistr	−r	swistrum	−rum

Gothic inflection classes, *-nd* declension

	friend			
	Singular		Plural	
Nom	frijōnds	−s	frijōnds	−s
Acc	frijōnd	−	frijōnds	−s
Gen	frijōndis	−is	frijōndē	−ē
Dat	frijōnd	−	frijōndam	−am

Gothic inflection classes

Looking at, e.g., the nominative and accusative singular forms of nouns from two inflection classes in Gothic, we see inverted lengths of the case markers:

(2) *-ja* declension (Gothic)

a. *har-jis* 'army-NOM'

b. *har-i* 'army-ACC'

(3) *-jō* declension (Gothic)

a. *band-i* 'band-NOM'

b. *band-ja* 'band-ACC'

→ Can the length differences be explained by the frequencies of the forms?

Testing the FFH:
The present approach

Corpora

We are using the Universal Dependencies Corpora (Silveira et al., 2014) for the following languages:

- Gothic (45k words)
- Latin (440k words)
- Ancient Greek (380k words)
- Modern Greek (51k words)
- Russian (1m words)
- Czech (1.3m words)
- Polish (72k words)
- Turkish (only used for comparison, 46k words)
- Latvian (only used for comparison, 44k words)
- Lithuanian (only used for comparison, 40k words)

We chose these languages based on their availability in the corpus, their case paradigms and inflection classes.

Inter-corpora comparison

A comparability problem

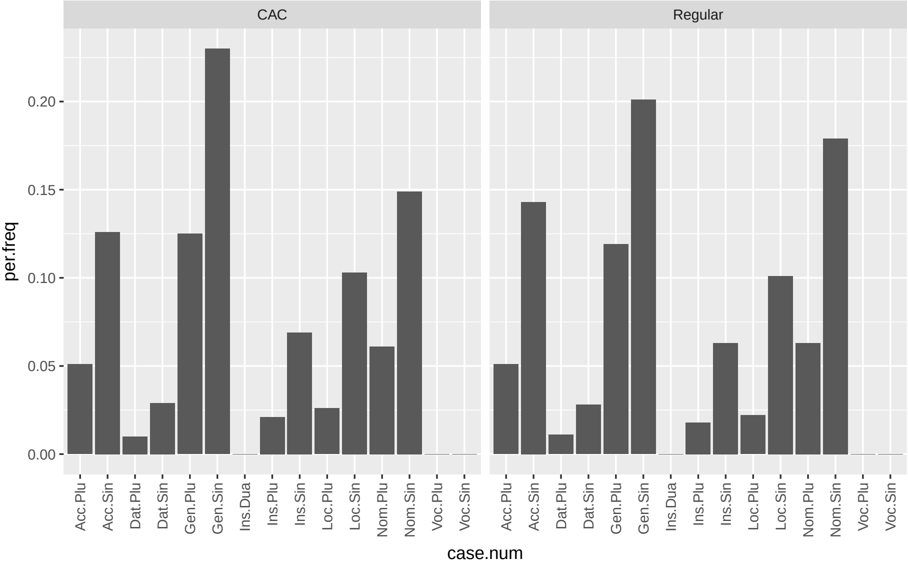
- As the corpora for the languages considered are not very homogeneous; we do not know to what extent the types of the corpora affect the use of different cases.

But: For Czech, Russian, Greek, and Latin, there are two different corpora available:

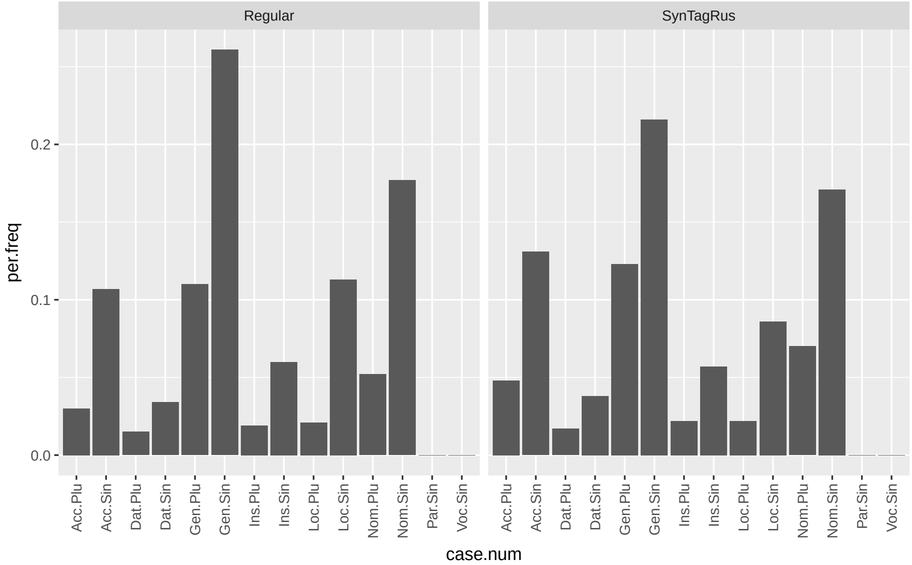
- Czech: Regular (1.3m words), CAC (482k words), overlap=0.01%
- Russian: Regular (87k words), SynTagRus (988k words), overlap=0%
- Greek: Regular (182k words), PROIEL (198k words), overlap=0.8%
- Latin: ITTB (280k words), PROIEL (159k words), overlap=0.006%

Therefore, we can compare the behaviour of different corpora for the same language in the UD data base with respect to case-number distribution.

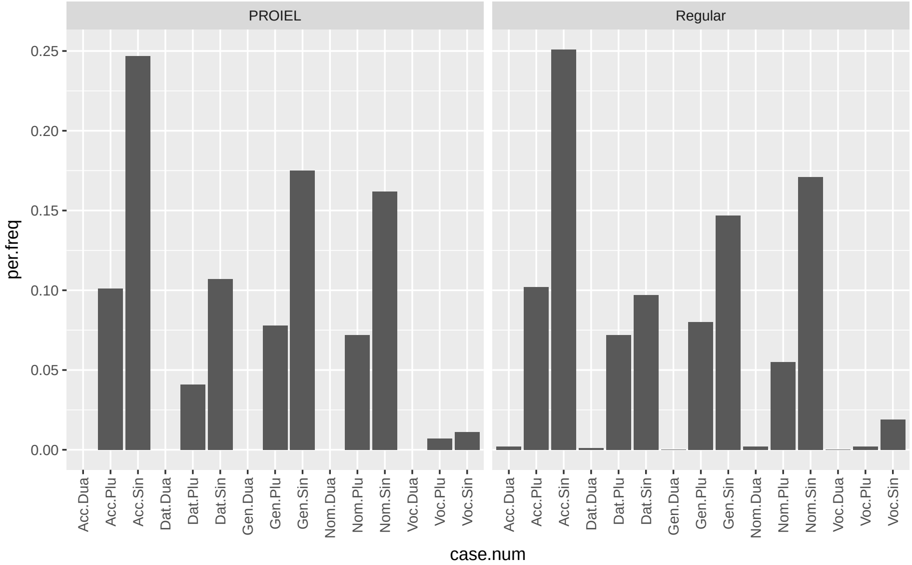
Czech


CAC
Regular
per.freq
0.00
0.05
0.10
0.15
0.20
Acc.Plu
Acc.Sin
Dat.Plu
Dat.Sin
Gen.Plu
Gen.Sin
Ins.Dua
Ins.Plu
Ins.Sin
Loc.Plu
Loc.Sin
Nom.Plu
Nom.Sin
Voc.Plu
Voc.Sin
case.num

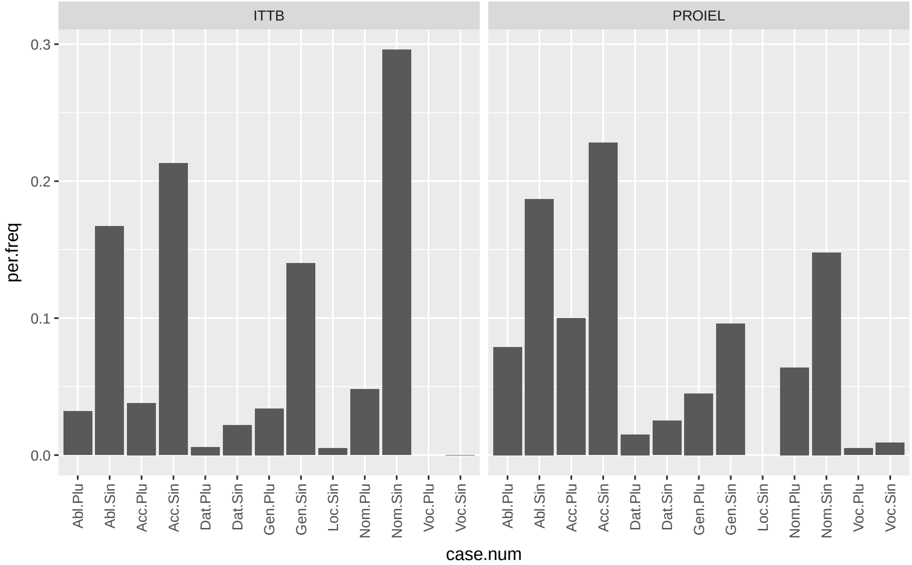
Russian


Regular
SynTagRus
per.freq
0.0
0.1
0.2
Acc.Plu
Acc.Sin
Dat.Plu
Dat.Sin
Gen.Plu
Gen.Sin
Ins.Plu
Ins.Sin
Loc.Plu
Loc.Sin
Nom.Plu
Nom.Sin
Par.Sin
Voc.Sin
case.num

Greek


PROIEL
Regular
per.freq
0.25
0.20
0.15
0.10
0.05
0.00
Acc.Dua
Acc.Plu
Acc.Sin
Dat.Dua
Dat.Plu
Dat.Sin
Gen.Dua
Gen.Plu
Gen.Sin
Nom.Dua
Nom.Plu
Nom.Sin
Voc.Dua
Voc.Plu
Voc.Sin
case.num

Latin


ITTB
PROIEL
0.3
0.2
0.1
0.0
per.freq
Abl.Plu
Abl.Sin
Acc.Plu
Acc.Sin
Dat.Plu
Dat.Sin
Gen.Plu
Gen.Sin
Loc.Sin
Nom.Plu
Nom.Sin
Voc.Plu
Voc.Sin
case.num

Are the case frequency distributions universal?

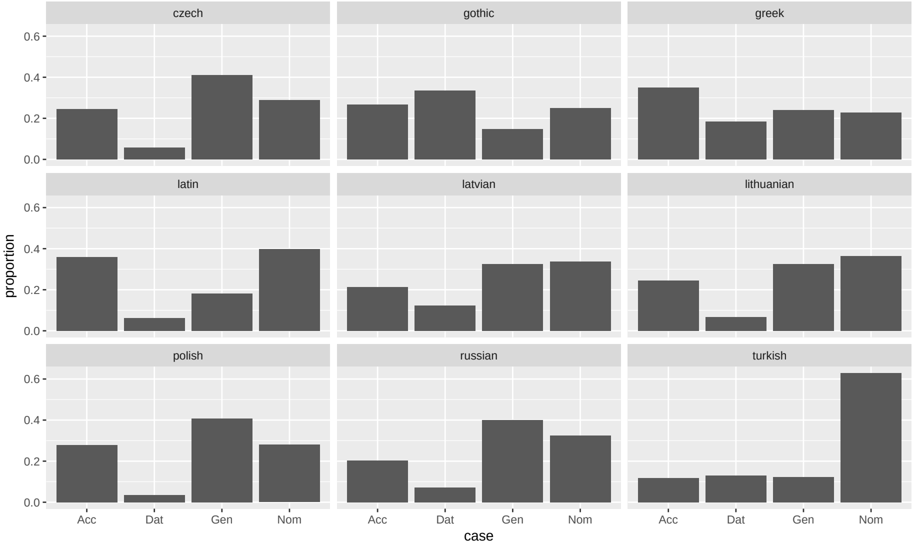
Case frequencies

Frequency has been used in previous studies to explain coding asymmetries of case markers

However, it does not yet seem to be clear whether case frequency is language specific or stable across languages (e.g. Downing and Stiebels, 2012, p. 402)

To address this issue, we looked at the frequencies of the nominative, accusative, dative, and genitive cases in different languages for nouns and pronouns.

Case frequencies



Testing the FFH

Testing the FFH

Is there a correlation between the frequency of a given case marker and its length? Do we find this effect across languages?

For simplicity and comparability, we only focus on the nominative, accusative, dative (except modern Greek), and genitive cases.

Length is measured as the number of segments/moras (if possible):

(4) *tugg-ōns* 'tongue' (GEN.SG, ACC.PL) (Gothic)

- ō $\rightarrow$ 2
- n $\rightarrow$ 1
- s $\rightarrow$ 1

Token / type frequency

We considered two kinds of frequency, token and type frequency.

In addition, the number of cells a case marker fills (i.e. the number of case values it expresses) has been noted.

The three values are calculated as follows:

(5) lex_1-CM (NOM.SG); lex_2-CM (NOM.SG);
lex_3-CM (NOM.SG); lex_2-CM (DAT.SG);
lex_1-CM (DAT.SG); lex_1-CM (NOM.SG)

Token frequency How often does a given marker occur in the corpus?
n_{token}(CM)=6

Type frequency With how many different lexemes does a given marker occur in the corpus?
n_{type}(CM)=3

Number of cases How many cells of the paradigm does a case marker fill?
n_{cases} (CM)=2

Cell distribution of the case marker

An example for the number of cases from Modern Greek:

(6) the distribution of the marker [-is]: (Modern Greek)

δυνάμ-εις	NOM.PL	'force'
δυνάμ-εις	ACC.PL	'force'
μάχ-ης	GEN.SG	'battle'

Results. Ancient Greek

Token frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.0000	0.1166	17.152	<2e-16	***
token.freq.scaled	-0.2502	0.1181	-2.118	0.0409	*
R-squared: 0.1082					

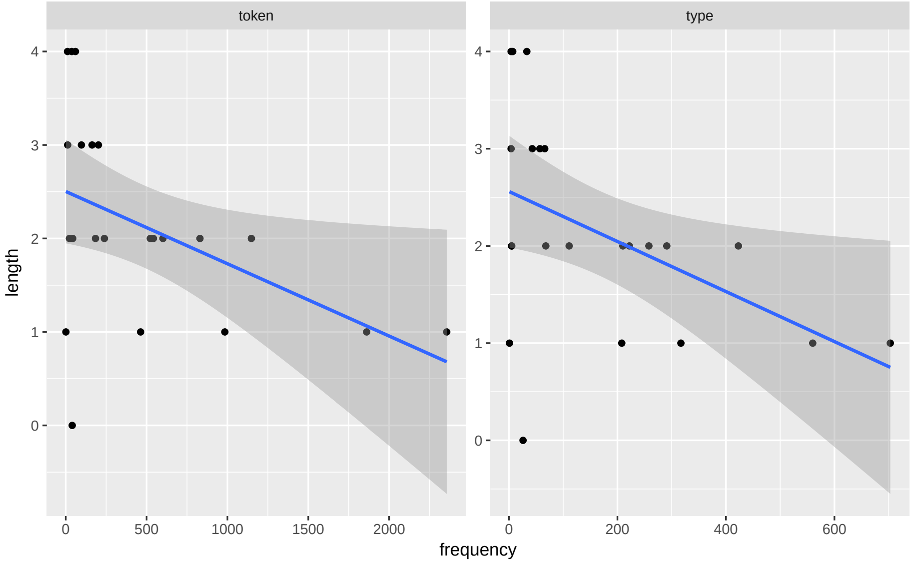
Type frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.0000	0.1171	17.086	<2e-16	***
type.freq.scaled	-0.2421	0.1186	-2.042	0.0483	*
R-squared: 0.1013					

Number of cases

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.0000	0.1146	17.446	<2e-16	***
ncases.numbers.scaled	-0.2826	0.1161	-2.434	0.0199	*
R-squared: 0.138					

Results. Ancient Greek



Results. Modern Greek

Token frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.1364	0.2110	10.127	2.56e-09	***
token.freq.scaled	-0.4884	0.2159	-2.262	0.035	*
R-squared: 0.2037					

Type frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.1364	0.2091	10.215	2.21e-09	***
type.freq.scaled	-0.5045	0.2141	-2.357	0.0287	*

R-squared: 0.2173

Results. Latin

Token frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.7436	0.1802	15.221	< 2e-16	***
token.freq.scaled	-0.6987	0.1826	-3.827	0.000484	***
R-squared: 0.2835					

Type frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.7436	0.1767	15.528	< 2e-16	***
type.freq.scaled	-0.6472	0.1790	-3.616	0.00091	***
type.res.ncases	-0.4598	0.2126	-2.162	0.03731	*
R-squared: 0.3302					

Results. Gothic

Token frequency

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.7273	0.1172	23.27	< 2e-16	***
token.freq.scaled	-0.5913	0.1183	-5.00	6.65e-06	***

Coefficients; R-squared: 0.3205

Type frequency

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.7273	0.1168	23.347	< 2e-16	***
type.freq.scaled	-0.5953	0.1179	-5.049	5.58e-06	***

Coefficients; R-squared: 0.3248

Results. Russian

Token frequency

<table>
<tr><th colspan="6">Coefficients</th></tr>
<tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(>|t|)</th><th></th></tr>
<tr><td>(Intercept)</td><td>1.6400</td><td>0.1314</td><td>12.484</td><td>9.99e-12</td><td>***</td></tr>
<tr><td>token.freq.scaled</td><td>-0.4932</td><td>0.1341</td><td>-3.679</td><td>0.00125</td><td>**</td></tr>
<tr><td colspan="6">R-squared: 0.3704</td></tr>
</table>

Type frequency

<table>
<tr><th colspan="6">Coefficients</th></tr>
<tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(>|t|)</th><th></th></tr>
<tr><td>(Intercept)</td><td>1.5775</td><td>0.3605</td><td>4.375</td><td>0.000293</td><td>***</td></tr>
<tr><td>type.freq.scaled</td><td>-0.5135</td><td>0.1284</td><td>-3.999</td><td>0.000705</td><td>***</td></tr>
<tr><td>(type.res.ncases, 3)1</td><td>1.6011</td><td>1.2647</td><td>1.266</td><td>0.220077</td><td></td></tr>
<tr><td>(type.res.ncases, 3)2</td><td>-2.1015</td><td>0.9180</td><td>-2.289</td><td>0.033074</td><td>*</td></tr>
<tr><td>(type.res.ncases, 3)3</td><td>0.3593</td><td>0.7403</td><td>0.485</td><td>0.632676</td><td></td></tr>
<tr><td colspan="6">R-squared: 0.5081</td></tr>
</table>

Results. Polish

Token frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.1000	0.1287	8.546	2.71e-05	***
token.freq.scaled	-0.4183	0.1357	-3.083	0.0151	*
R-squared: 0.5429					

Type frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.1000	0.1272	8.651	2.48e-05	***
type.freq.scaled	-0.4225	0.1340	-3.152	0.0136	*
R-squared: 0.554					

Results. Czech

Token frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.7600	0.1236	14.234	1.40e-12	***
token.freq.scaled	-0.6458	0.1262	-5.118	3.96e-05	***
token.res.ncases	-0.3959	0.1673	-2.366	0.0272	*
R-squared: 0.591					

Type frequency

Coefficients					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.7600	0.1224	14.377	1.15e-12	***
type.freq.scaled	-0.6561	0.1249	-5.251	2.87e-05	***
type.res.ncases	-0.3830	0.1664	-2.302	0.0312	*
R-squared: 0.5991					

To sum up

Case frequency

- Different languages exhibit different case frequency distributions.

The Form-Frequency Hypothesis

- Raw token frequency had some explanatory power for the length of the exponents.
- Type frequency and the number of cases were better predictors.
- However, frequency alone does not explain the asymmetry of the length of case markers.

→ What other factors can play a role that are able to outrank the frequency effects?