Quantitative approaches to analysing sign language corpus data



Adam Schembri

with Jordan Fenlon & Kearsy Cormier

Other collaborators: Bencie Woll, Ceil Lucas, David McKee, Della Goswell, Donovan Cresdee, Rachel McKee, Ramas Rentelis, Robert Bayley, Rose Stamp, Sara Pivac & Trevor Johnston

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Australian Government
Australian Research Council

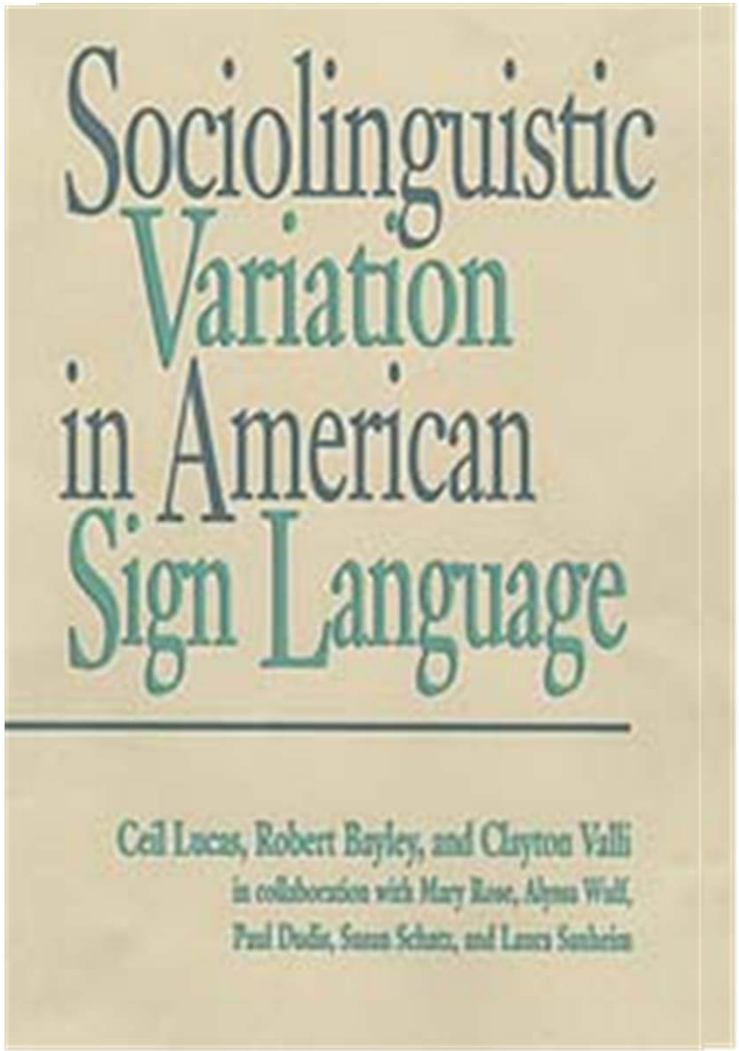
LA TROBE UNIVERSITY

UCL

Variationist approaches to sign language research

- Since his pioneering work on sociolinguistic variation and change in American English in the 1960s, William Labov has led a tradition of language studies with a focus on variation.
- Labov's notion of the 'linguistic variable' refers to two or more variant forms in a language that vary according to linguistic and social factors.
- This approach has come to be known as *variationist linguistics*
- This approach has always incorporated a statistical component that is used to describe patterns of variation between alternative forms in large datasets of naturalistic language use.

Variationist approaches to sign language research



- Work within variationist (socio-)linguistics often has implications for linguistic theory in general, as the quantitative approach to the study of corpora can be used to test any linguistic theory.
- The application of variationist quantitative approaches to the study of sign languages was pioneered by Ceil Lucas and her colleagues in the early 1990s on American Sign Language (ASL).
- Over the last decade, have grown to include work on sign languages in Australia, New Zealand, Italy and the United Kingdom.

Variationist approaches to sign language research

- In this presentation, I give an overview of some recent studies that have involved drawing on sign language corpora.
- Quantitative analysis has shown that the factors that drive sociolinguistic variation and change in both spoken and signed language communities appear to be broadly similar.
- It has also demonstrated that some factors involved in variation in sign languages are distinctive.

Variationist approaches to sign language research

- In this talk, I will discuss how the data in such studies have been coded for quantitative analysis, and how the coding is analysed using Rbrul software, a multivariate statistical package designed specifically for (socio-)linguistic studies of large language samples.
- I will then discuss how to interpret the statistical results, and what conclusions can be drawn about the nature of variation and change in these studies.
- For this section, I will focus on work on grammatical variation in BSL (British Sign Language).

Varationist approaches to sign language research

- In this talk, I will discuss how the data in such studies have been coded for quantitative analysis, and how the coding is analysed using Rbrul software, a multivariate statistical package designed specifically for (socio-)linguistic studies of large language samples.
- I will then discuss how to interpret the statistical results, and what conclusions can be drawn about the nature of variation and change in these studies.
- For this section, I will focus on work on grammatical variation in BSL (British Sign Language).

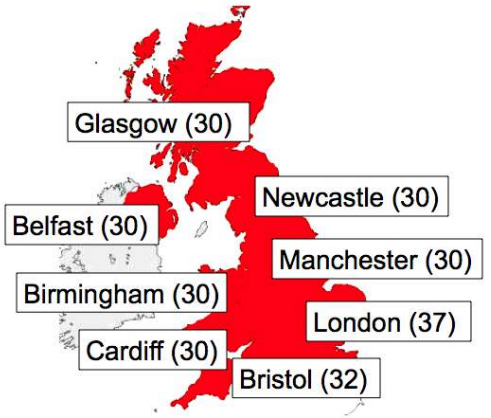
BANZSL projects

- Sociolinguistic variation in Auslan project: 2003-2005
  (researcher with Trevor Johnston & Della Goswell)
- Auslan Corpus project: 2004-present
  (researcher with Trevor Johnston)
- Sociolinguistic variation in NZSL project
  (consultant for David & Rachel McKee)
- BSL Corpus Project
  (project director, working with Jordan Fenlon, Ramas Rentelis, Rose Stamp & Kearsy Cormier)



Background to the projects

- Auslan: 2 projects filmed 256 Deaf participants in 5 cities across Australia
- NZSL: 138 Deaf participants from 5 cities and towns across New Zealand
- BSL: 249 Deaf participants in 8 cities in the UK
- Projects used elicited narratives, spontaneous narratives, free conversation, interviews, lexical elicitation, responses to video stimuli, barrier games



Studies to date

- Phonological variation and change
  - Location variation in Auslan & NZSL (Schembri et al., 2009)
  - Handshape and orientation variation in BSL (Fenlon et al., 2013)
- Lexical variation and change
  - Number signs in NZSL (McKee et al., 2010)
  - Number, color and place name signs in BSL (Stamp et al., 2014)
  - Fingerspelling in Auslan (Schembri & Johnston, 2007)
  - Mouth actions in Auslan (Johnston et al., 2014)
- Grammatical variation and change
  - Variable subject expression in Auslan & NZSL (McKee et al., 2011)
  - Indicating verbs in Auslan (de Beuzeville et al., 2009), and BS L (Fenlon et al., 2014)
  - Perfective aspect marking in Auslan (Johnston et al., submitted)

Directional verbs

- Directional verbs: a class of verbs that move in space between locations associated with ‘subjects’ and ‘objects’
  - Occur in vast majority of sign languages studied to date
- In this study, we focus on a subset of directional verbs that denote a transfer between animate and inanimate arguments (as opposed to marking location)

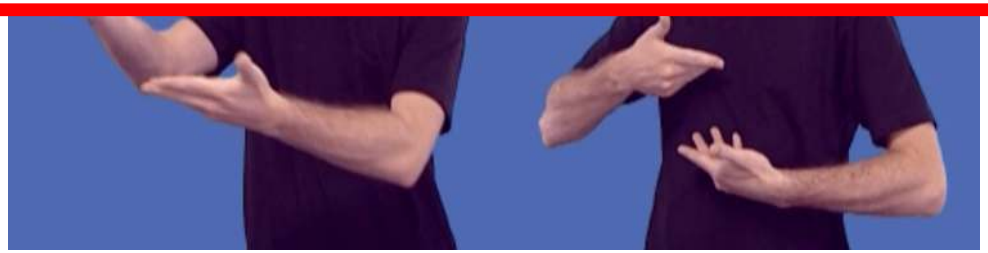
BSL GIVE ‘She gives me’

Directional verbs

- Directional verbs: a class of verbs that move in space between locations associated with ‘subjects’ and ‘objects’
  - Occu…te
- In this s…erbs that denote … argume…

How to analyse these verbs?

Two views



BSL GIVE ‘She gives me’

1. Agreement

- These verbs show grammatical agreement with subject/object (Padden 1988, Lillo-Martin & Meier 2011) cf. Spanish

1st person	***Yo* hablo "I speak"**
2nd person	*Tu* hablas "you speak"
3rd person	*El/Ella* habla 'he/she speaks'

- Object agreement is claimed to be obligatory (Lillo-Martin & Meier, 2011)

2. Fusion of verbs and gesture

- These verbs point to their present referents or locations associated with their imagined referents (Liddell, 2003, de Beuzeville et al. 2009)
  - Appear to be affected by language external factors
  - Not obligatory

Which analysis?

- Some linguists accept the gestural argument whilst still arguing that modification reflects grammatical agreement (e.g., Lillo-Martin & Meier; 2011, Mathur & Rathmann, 2010)

- Some issues
  - Different theoretical assumptions about the relationship between language and other aspects of communication such as gesture
  - Lack of usage data (although see de Beuzeville et al., 2009)

BSL Corpus

1680 tokens of directional verbs in free conversation from 100 signers (4 regions) in the BSL Corpus (Schembri et al., 2010)

Coded for

- Coded for actor and undergoer modification according to:
  - Person, number, animacy and co-reference
  - Presence/absence of constructed action
  - Direction of movement
  - Lexical frequency (using objective frequency measures from Fenlon et al., 2014)
  - Social factors: gender, age, region, language background, ethnicity
- Included lexical items and participants as a random effect in a mixed effect model

Verbs in our data

- 93 verbs
- Top 10 verbs account for 56% of 1680 tokens
  - SAY, LOOK, LOOK2, MEET, GIVE, PAY, DISCUSS, GIVE-INFORMATION, ASK

Rate of modification

Undergoer

UNMODIFIED

Does not differ from citation form (typically first to second person)

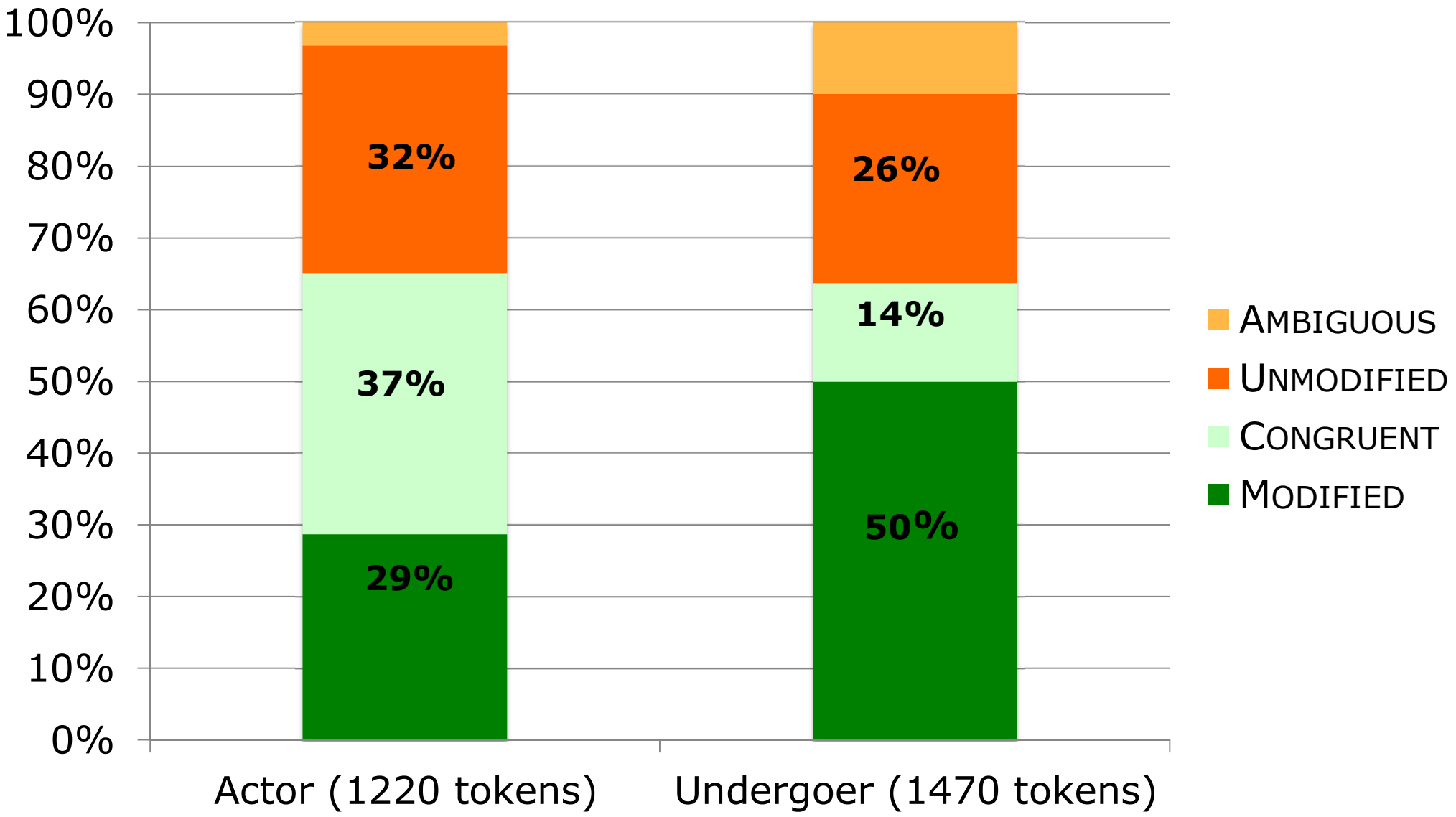
MODIFIED

differs from citation form

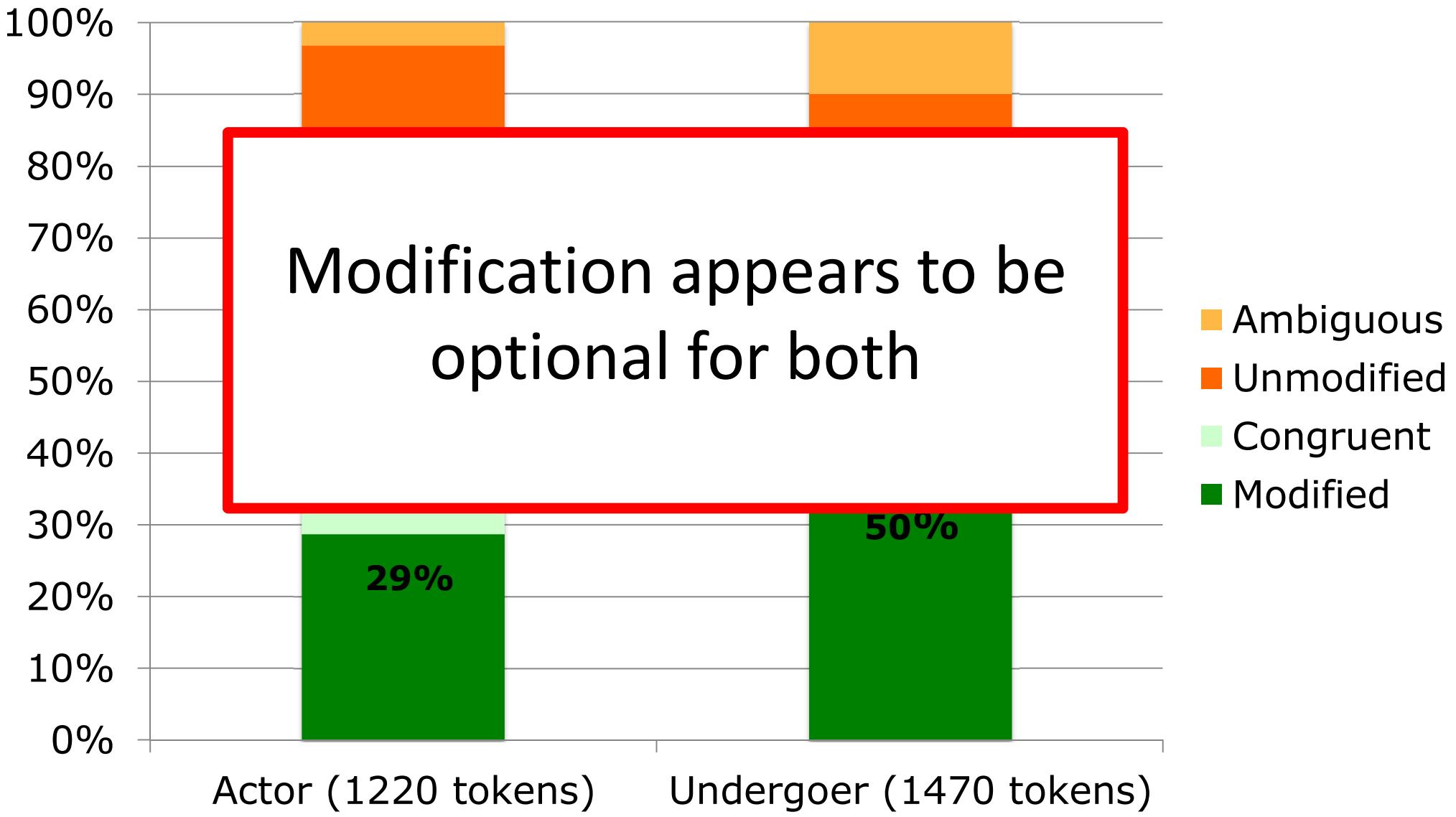
CONGRUENT

looks like citation form but matches spatial location of referent

Rate of modification



Rate of modification



Factors underlying actor modification

- Only person is significant
  - First person strongly favours modification over third and second person referents (93% over 54%, 52% respectively, p=<.001)

Factors underlying undergoer modification

- **Coreference**
  - Coreference with a noun, null argument, pronoun occurs with modification more than no coreference (p=<.001)
    - *Suggests a reference tracking system*
- **Constructed action**
  - Overt CA is more likely to occur with modification than no CA (p= <.001)
    - *Suggests that signers are pointing to imagined referents*

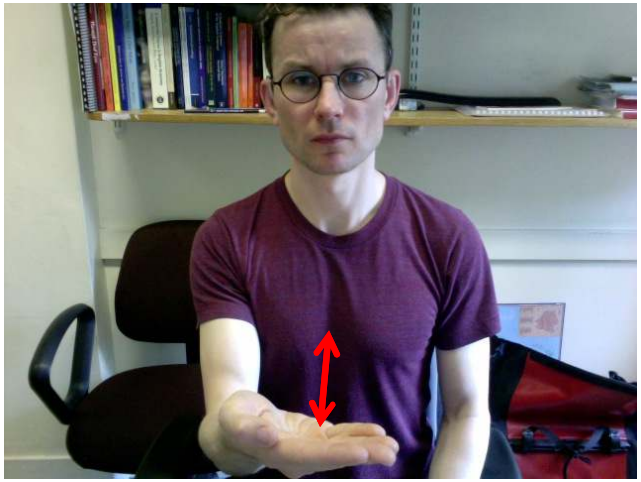
Factors underlying undergoer modification

- **Person**
  - Second and first person occurs with modification more than third person (p =<.01)
    - *Reflects a distinction between present and non-present referents*

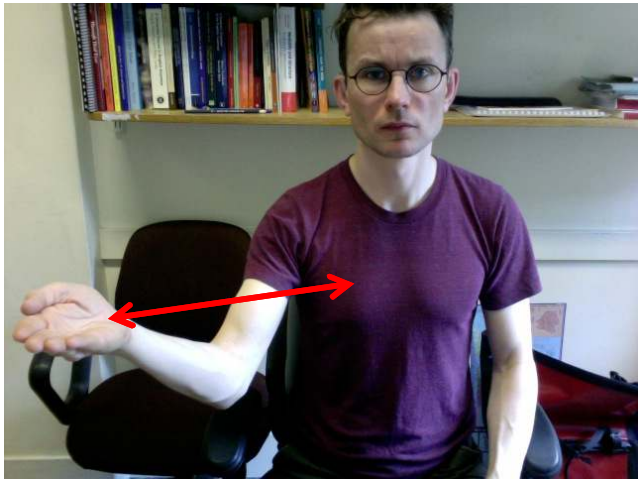
- **Animacy**
  - Animate arguments occurs with more modification than inanimate arguments (p=<.01)
    - *Is linked to person because inanimate referents are likely to be in third person*

Direction of movement

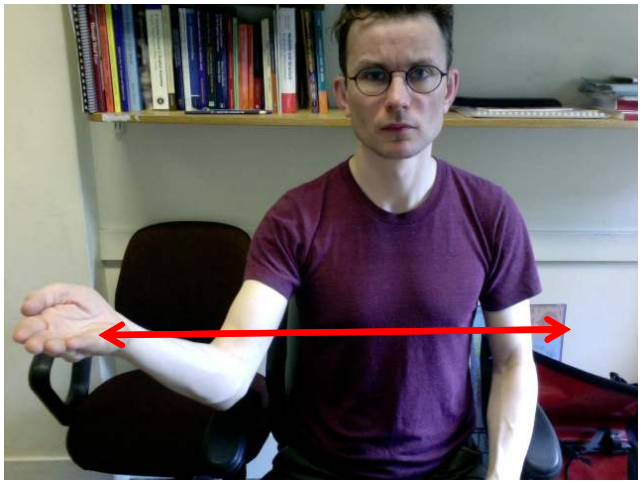
For constructions using third to third person referents (*John gave Mary a book*) – which axis do signers use?



Z axis

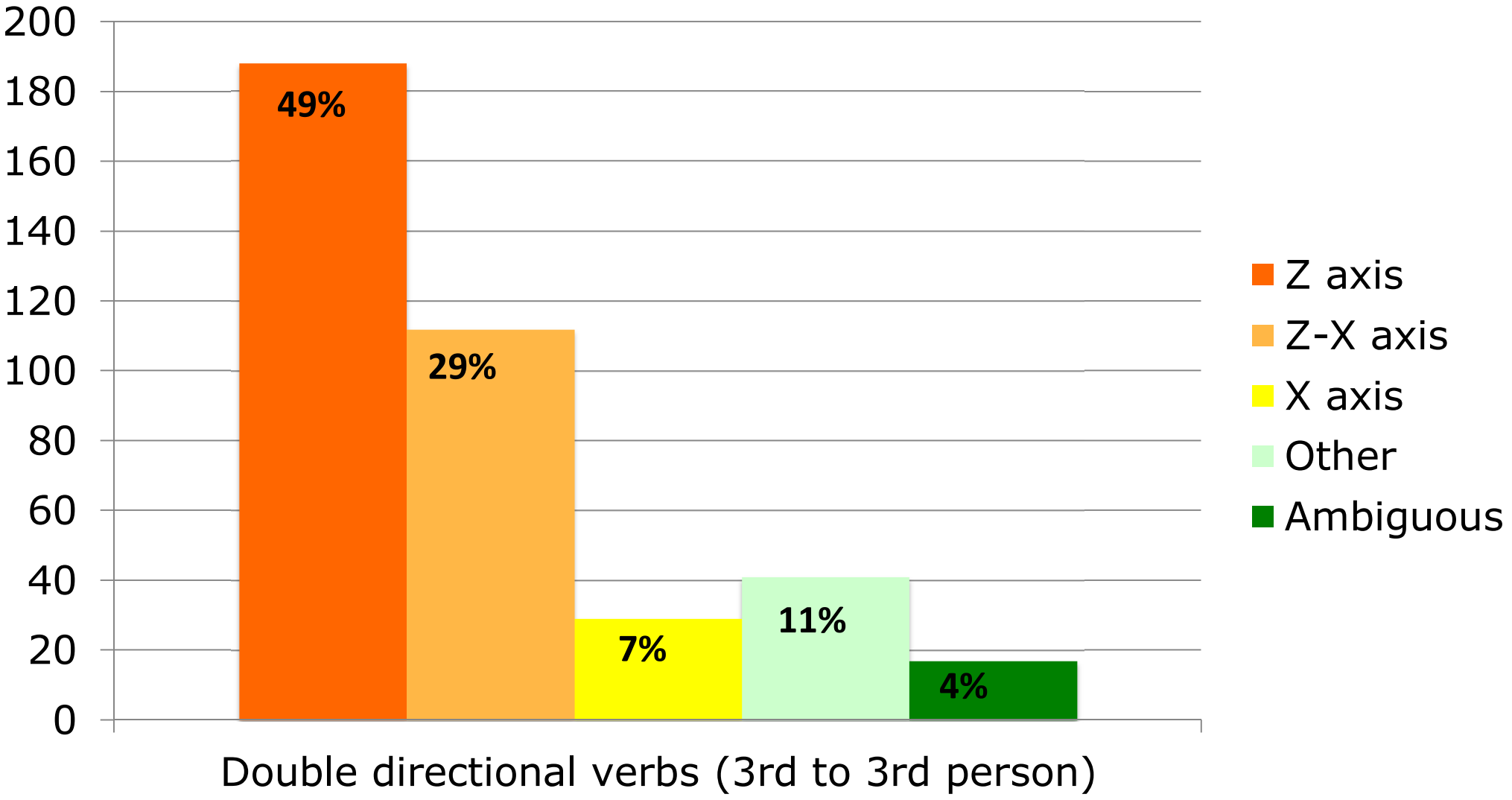


Z-X axis

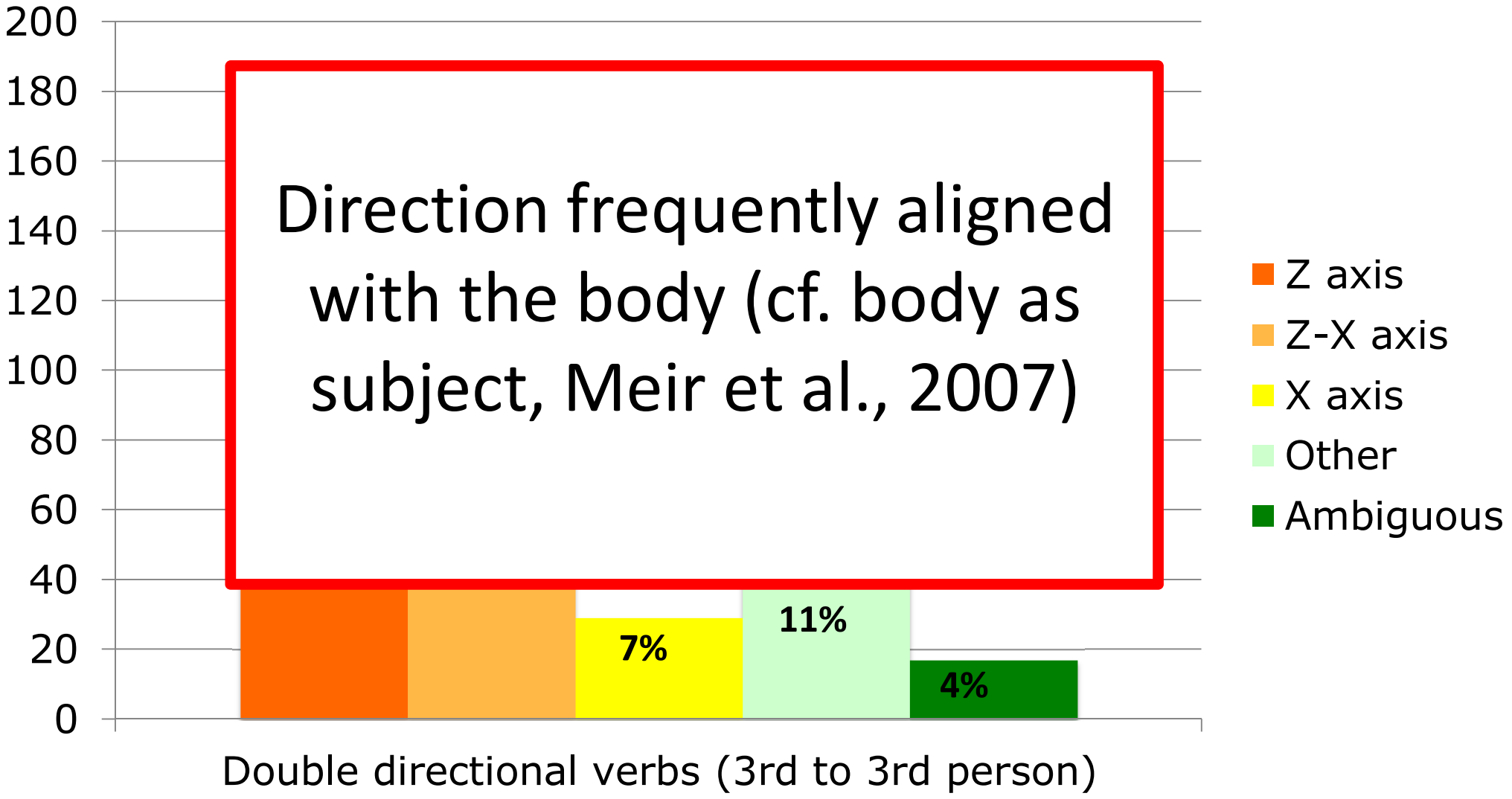


X axis

Direction of movement



Direction of movement



Non significant factors

- Frequency is not significant
- Social factors are not significant
- Lack of finding regarding age and frequency suggests there is not a change in progress
  - i.e., *little evidence of grammaticalisation in BSL with respect to this subset of directional verbs*

Conclusion

- Preliminary findings from this study seem to support the idea of directionality and modification as a pointing-based reference tracking system
  - Does not seem to support the agreement analysis of these verbs

Factors underlying actor modification

Factor	Tokens	% modified	Log odds
Person (p=<.001)			
First	334	93%	1.974
Third	432	54%	-0.900
Second	85	52%	-1.074

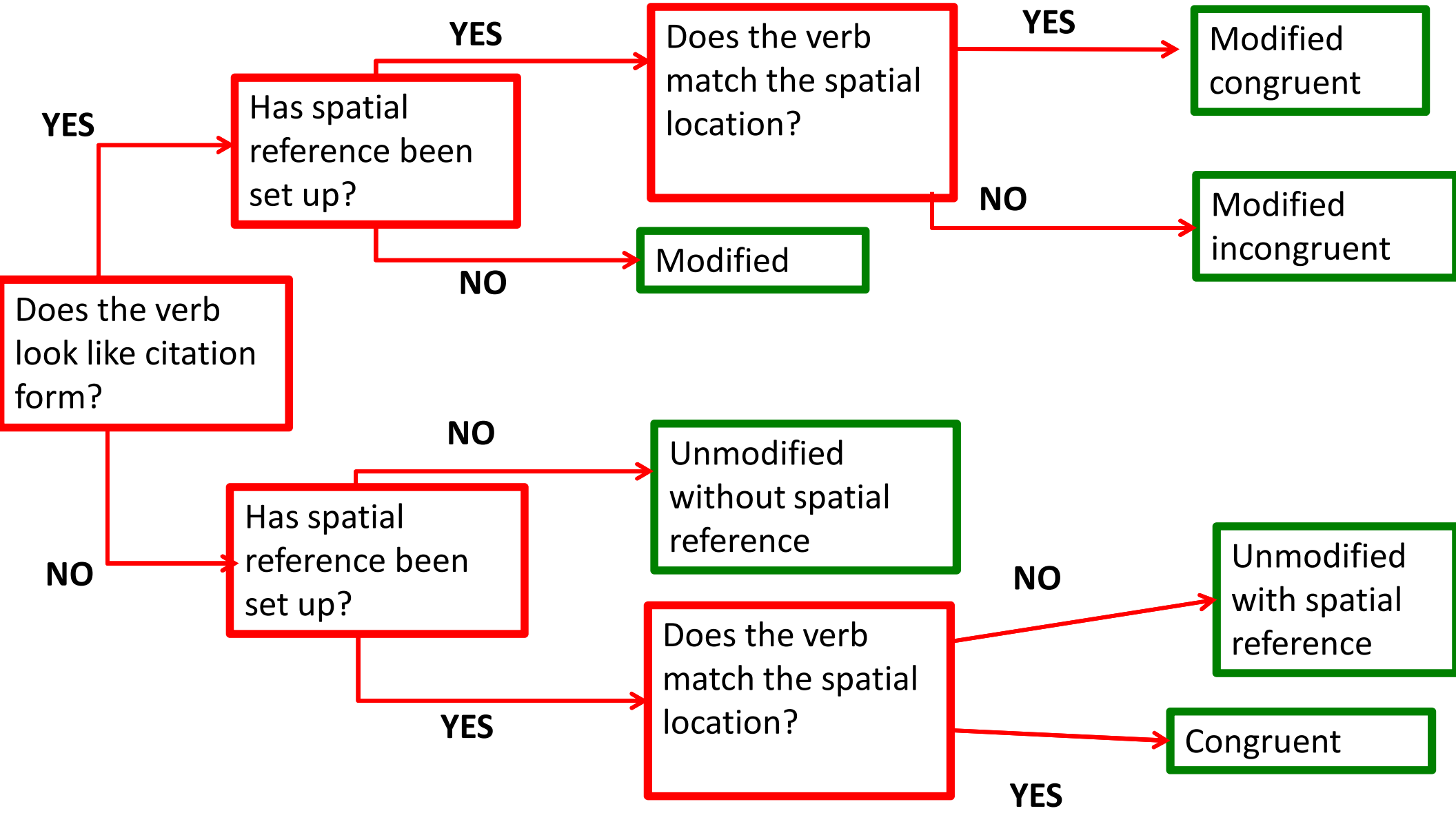
Factors underlying undergoer modification

Factor	Tokens	% modified	Log odds
Coreference (p=<.001)			
Noun	81	74%	0.460
Null argument	266	76%	0.134
Pronoun	126	73%	-0.086
No coreference	767	60%	-0.508
Constructed Action (p=<.001)			
CA with eyegaze and other articulators	504	73%	0.329
CA without eyegaze	18	76%	0.272
Eyegaze only	351	62%	-0.222
No CA	367	59%	-0.379

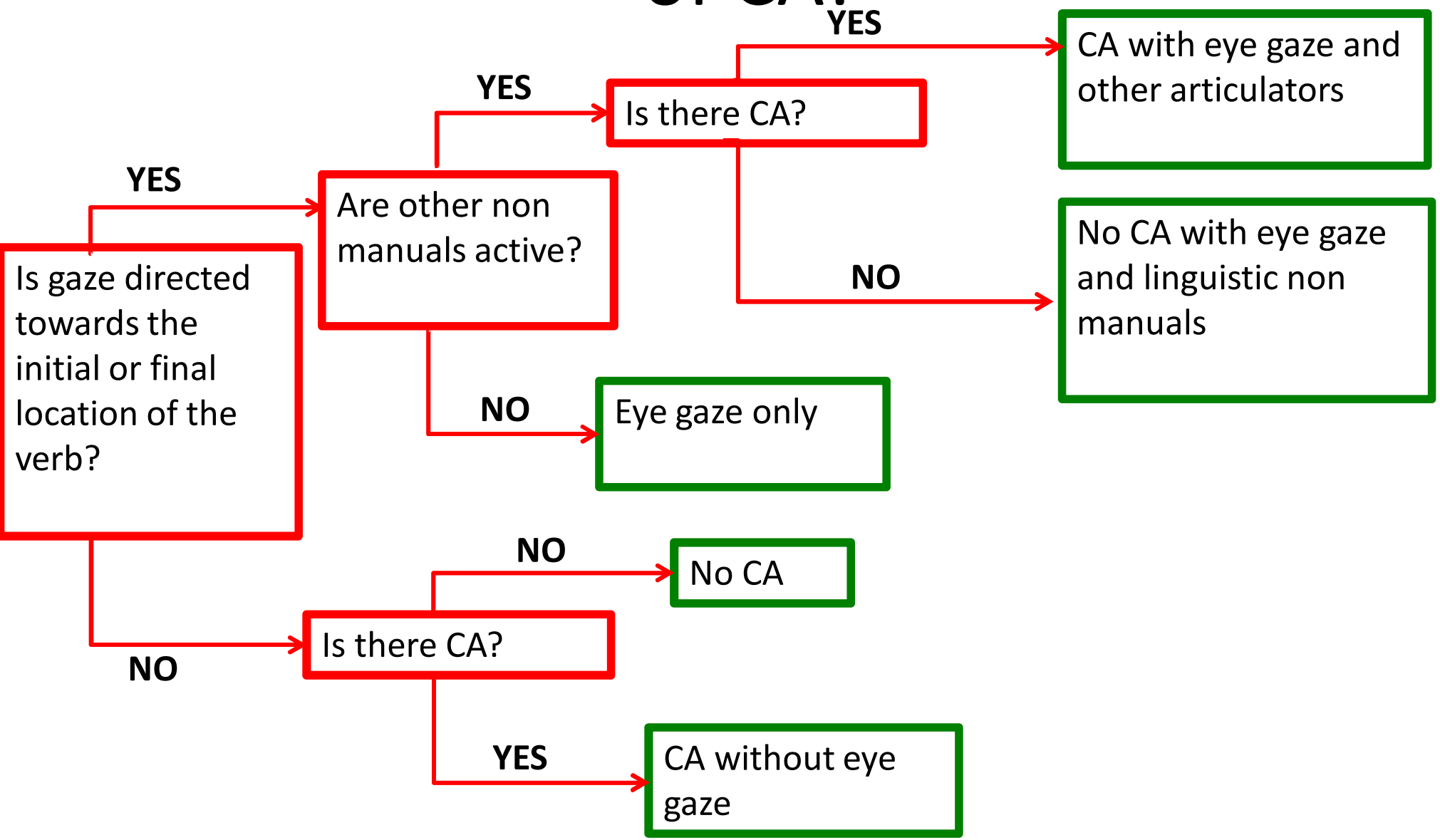
Factors underlying undergoer modification

Factor	Tokens	% modified	Log odds
Person (p =<.01)			
Second	59	76%	0.484
First	276	72%	0.003
Third	905	63%	-0.486
Animacy (p =<.01)			
Animate	754	70%	0.221
Ambiguous	130	71%	0.200
Other animates	129	66%	0.067
Non locative inanimates	227	51%	-0.488

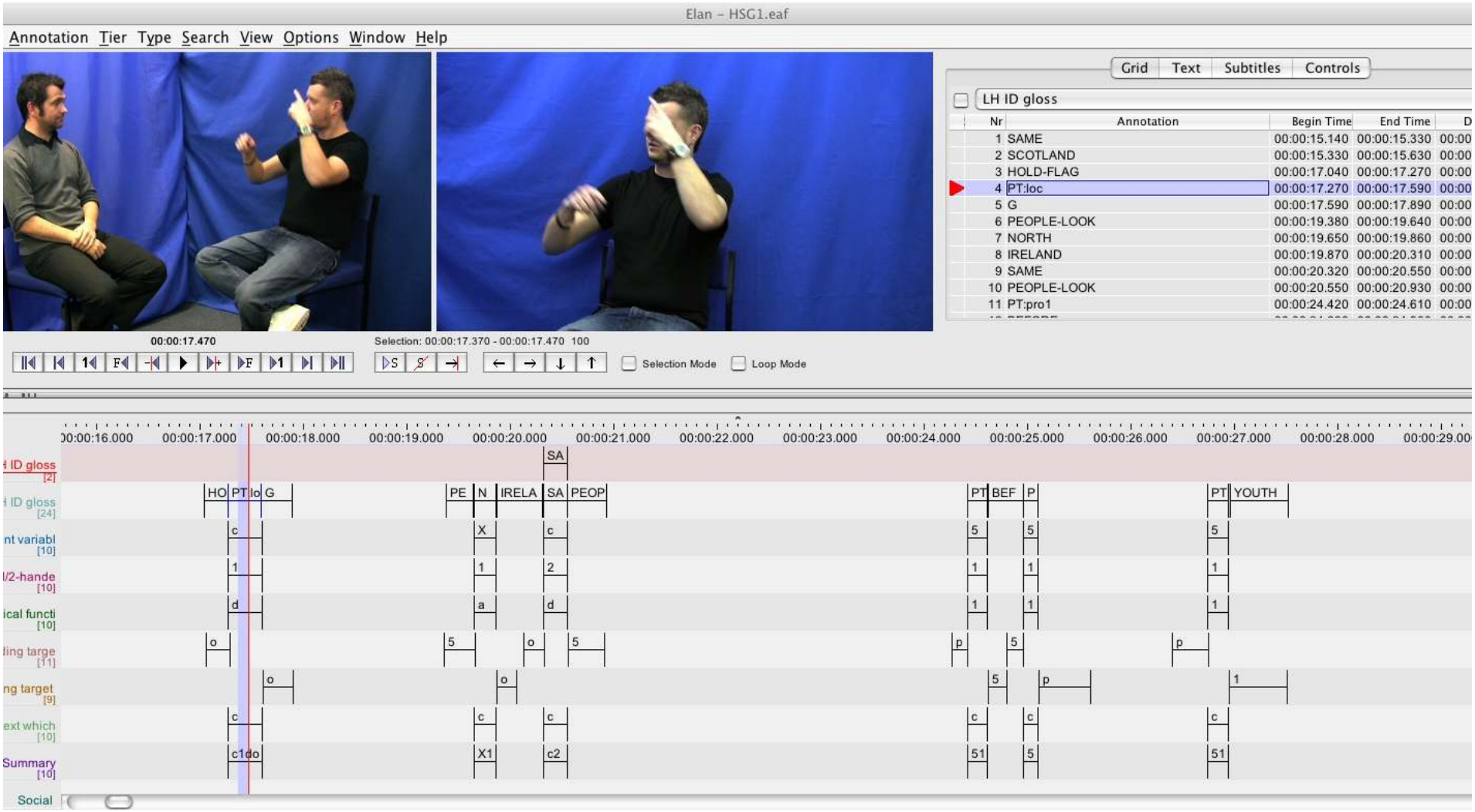
How did we decide modification?



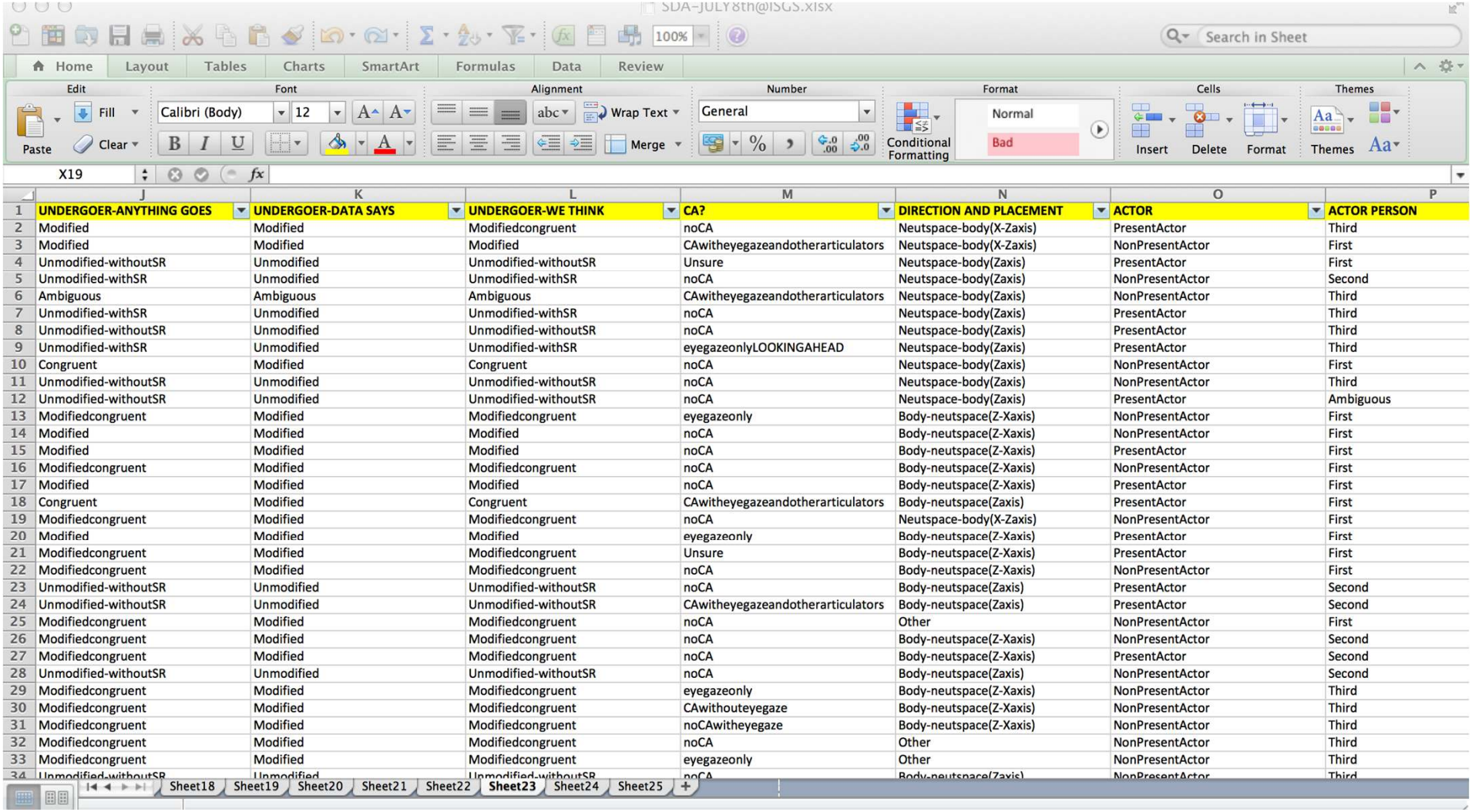
How did we decide presence/absence of CA?



Step 1: Annotations in ELAN



Step 2: Export to Excel



	J	K	L	M	N	O	P
1	UNDERGOER-ANYTHING GOES	UNDERGOER-DATA SAYS	UNDERGOER-WE THINK	CA?	DIRECTION AND PLACEMENT	ACTOR	ACTOR PERSON
2	Modified	Modified	Modifiedcongruent	noCA	Neutspace-body(X-Zaxis)	PresentActor	Third
3	Modified	Modified	Modified	CAwitheyegazeandotherarticulators	Neutspace-body(X-Zaxis)	NonPresentActor	First
4	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	Unsure	Neutspace-body(Zaxis)	PresentActor	First
5	Unmodified-withSR	Unmodified	Unmodified-withSR	noCA	Neutspace-body(Zaxis)	NonPresentActor	Second
6	Ambiguous	Ambiguous	Ambiguous	CAwitheyegazeandotherarticulators	Neutspace-body(Zaxis)	NonPresentActor	Third
7	Unmodified-withSR	Unmodified	Unmodified-withSR	noCA	Neutspace-body(Zaxis)	PresentActor	Third
8	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	noCA	Neutspace-body(Zaxis)	PresentActor	Third
9	Unmodified-withSR	Unmodified	Unmodified-withSR	eyegazeonlyLOOKINGAHEAD	Neutspace-body(Zaxis)	PresentActor	Third
10	Congruent	Modified	Congruent	noCA	Neutspace-body(Zaxis)	NonPresentActor	First
11	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	noCA	Neutspace-body(Zaxis)	NonPresentActor	Third
12	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	noCA	Neutspace-body(Zaxis)	PresentActor	Ambiguous
13	Modifiedcongruent	Modified	Modifiedcongruent	eyegazeonly	Body-neutspace(Z-Xaxis)	NonPresentActor	First
14	Modified	Modified	Modified	noCA	Body-neutspace(Z-Xaxis)	NonPresentActor	First
15	Modified	Modified	Modified	noCA	Body-neutspace(Z-Xaxis)	PresentActor	First
16	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Body-neutspace(Z-Xaxis)	NonPresentActor	First
17	Modified	Modified	Modified	noCA	Body-neutspace(Z-Xaxis)	PresentActor	First
18	Congruent	Modified	Congruent	CAwitheyegazeandotherarticulators	Body-neutspace(Zaxis)	PresentActor	First
19	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Neutspace-body(X-Zaxis)	NonPresentActor	First
20	Modified	Modified	Modified	eyegazeonly	Body-neutspace(Z-Xaxis)	PresentActor	First
21	Modifiedcongruent	Modified	Modifiedcongruent	Unsure	Body-neutspace(Z-Xaxis)	PresentActor	First
22	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Body-neutspace(Z-Xaxis)	NonPresentActor	First
23	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	noCA	Body-neutspace(Zaxis)	PresentActor	Second
24	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	CAwitheyegazeandotherarticulators	Body-neutspace(Zaxis)	PresentActor	Second
25	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Other	NonPresentActor	First
26	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Body-neutspace(Z-Xaxis)	NonPresentActor	Second
27	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Body-neutspace(Z-Xaxis)	PresentActor	Second
28	Unmodified-withoutSR	Unmodified	Unmodified-withoutSR	noCA	Body-neutspace(Zaxis)	NonPresentActor	Second
29	Modifiedcongruent	Modified	Modifiedcongruent	eyegazeonly	Body-neutspace(Z-Xaxis)	NonPresentActor	Third
30	Modifiedcongruent	Modified	Modifiedcongruent	CAwithouteyegaze	Body-neutspace(Z-Xaxis)	NonPresentActor	Third
31	Modifiedcongruent	Modified	Modifiedcongruent	noCAwitheyegaze	Body-neutspace(Z-Xaxis)	NonPresentActor	Third
32	Modifiedcongruent	Modified	Modifiedcongruent	noCA	Other	NonPresentActor	Third
33	Modifiedcongruent	Modified	Modifiedcongruent	eyegazeonly	Other	NonPresentActor	Third

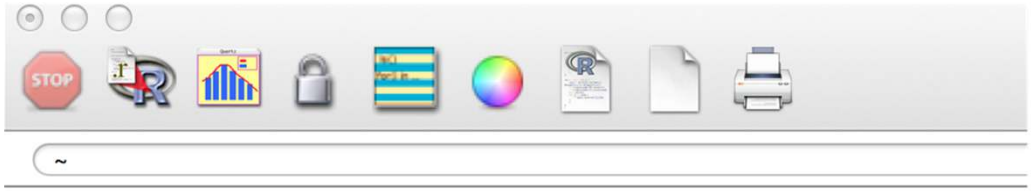
Step 3: check data

- Once exported to Excel, the data can be checked for inconsistencies.
- Rbrul will treat subtle differences in codes (e.g., differences in upper or lower case) as representing different factors, so consistency is important.
- Once checked, file must be saved as a text file.

Step 3: check data

- Every row of the spreadsheet needs to be a token (also known as an observation – a single instance of your dependent variable).
- One column must contain the dependent variable, or response.
- The rest of the columns must contain the independent variables, or predictors.
- You cannot have a spreadsheet where each row represents a single speaker and multiple tokens/observations from that speaker are in separate columns.

Step 4: Set up RBrul

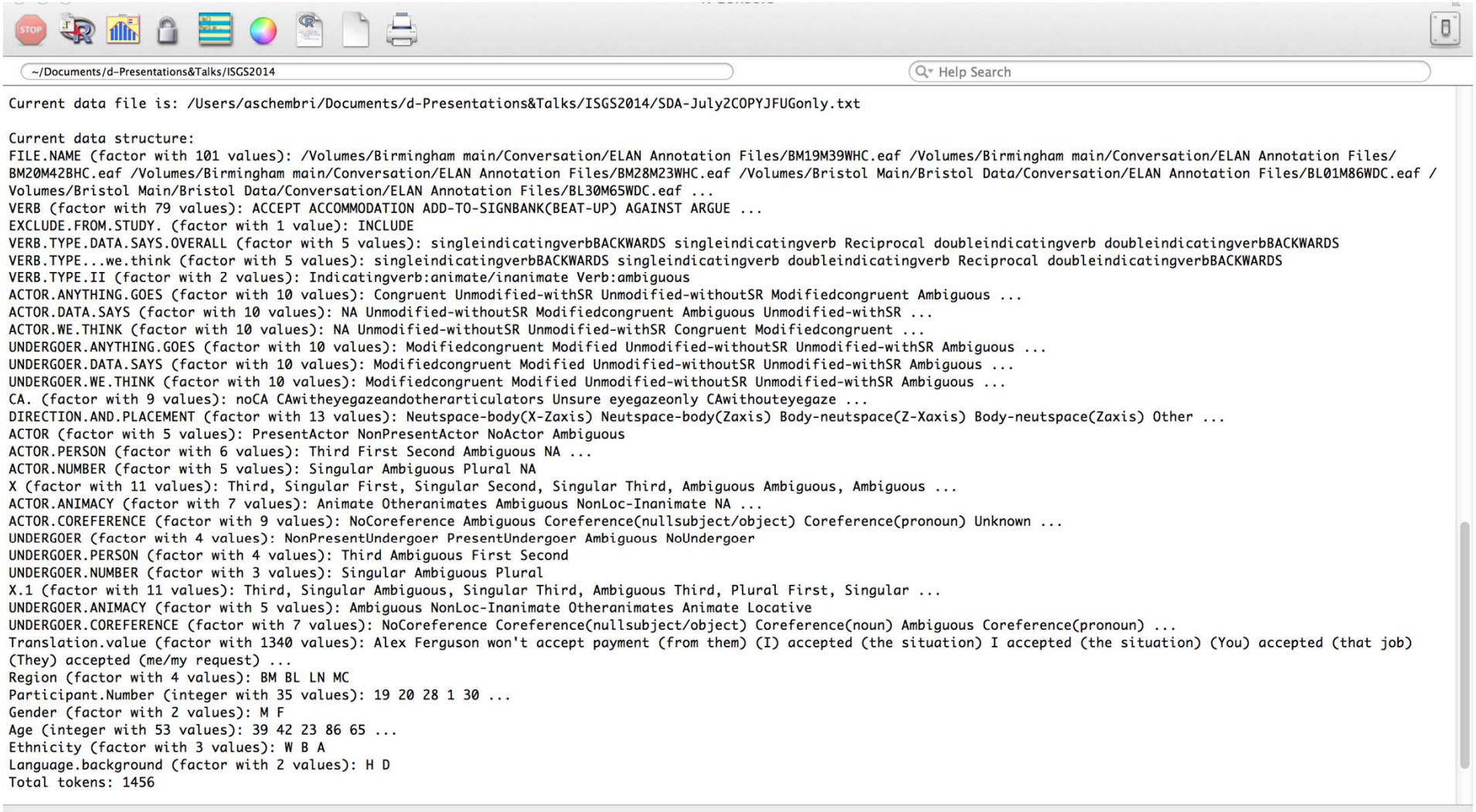



```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

[History restored from /Users/aschembri/.Rapp.history]

> source("http://www.danielezrajohnson.com/Rbrul.R")
```

Why Rbrul?

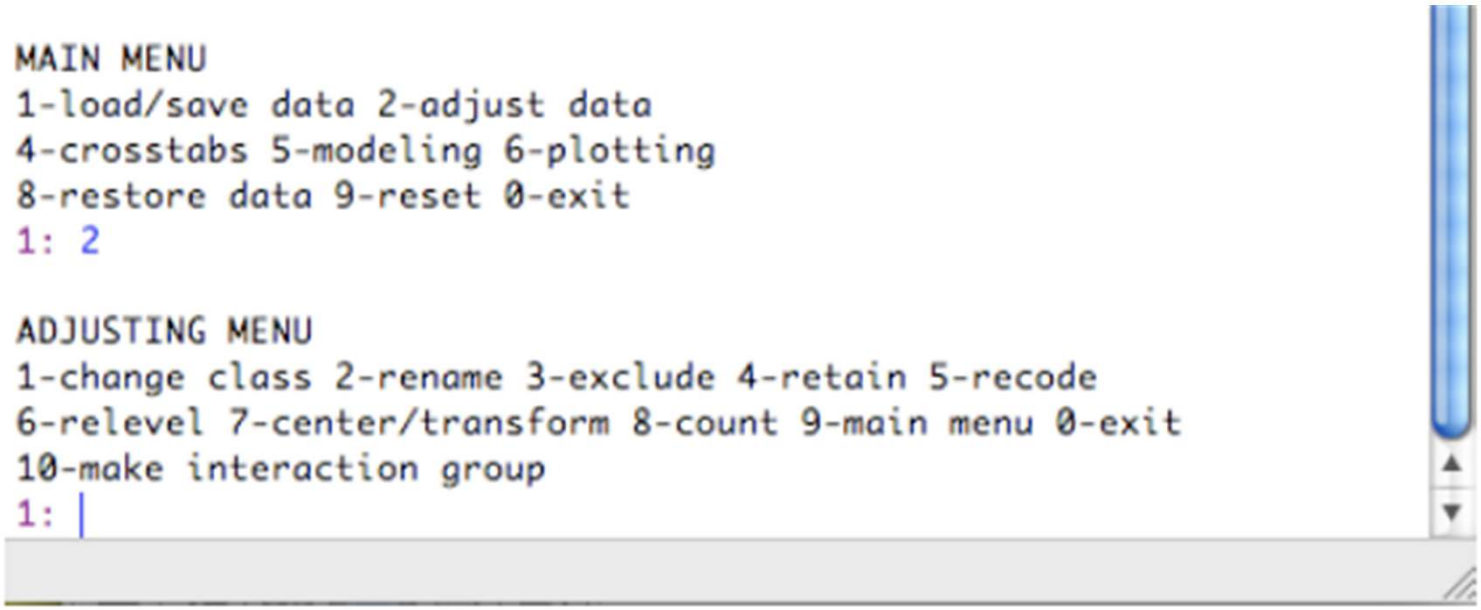
- Rbrul is a program specifically designed for analysing linguistic data within R
- Inspired by David Sankoff's original VARBRUL 'variable rule' program, and its successor Goldvarb which are widely used in variationist linguistics
- Rbrul carries out multiple regression analsyes with unbalanced binary data
- The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent variable.
- Rbrul is superior to Goldvarb because it supports both continuous (i.e., numerical) and categorical vraiables
- It also uses 'mixed effect models' with both random and fixed effects: this allows it to take into account by-speaker and by-item correlations

Step 5: Load the data to Rbrul



Step 6: Recode/exclude

- You can adjust the data by recoding (regrouping factors in each factor group) or by excluding factors



Step 7: Choose dependent variable

```
MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1: 5

No variables chosen.

MODELING MENU
1-choose variables 2-one-level (recommended)
3-step-up 4-step-down 5-step-up/step-down
6-trim 7-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-FILE.NAME 2-VERB 4-
VERB.TYPE.DATA.SAYS.OVERALL 5-VERB.TYPE...we.think 6-VERB.TYPE.II 7-ACTOR.ANYTHING.GOES 8-
ACTOR.DATA.SAYS 9-ACTOR.WE.THINK 10-UNDERGOER.ANYTHING.GOES 11-UNDERGOER.DATA.SAYS 12-
UNDERGOER.WE.THINK 13-CA. 14-DIRECTION.AND.PLACEMENT 15-ACTOR 16-ACTOR.PERSON 17-
ACTOR.NUMBER 18-X 19-ACTOR.ANIMACY 20-ACTOR.COREFERENCE 21-UNDERGOER 22-UNDERGOER.PERSON 23-
UNDERGOER.NUMBER 24-X.1 25-UNDERGOER.ANIMACY 26-UNDERGOER.COREFERENCE 27-Translation.value
28-Region 29-Participant.Number 30-Gender 31-Age 32-Ethnicity 33-Language.background)
1:
```

Step 8: Choose independent variables

```
1:
Choose predictors (independent variables) by number (1-FILE.NAME 2-VERB 4-
VERB.TYPE.DATA.SAYS.OVERALL 5-VERB.TYPE...we.think 6-VERB.TYPE.II 7-ACTOR.ANYTHING.GOES 8-
ACTOR.DATA.SAYS 9-ACTOR.WE.THINK 10-UNDERGOER.ANYTHING.GOES 12-UNDERGOER.WE.THINK 13-CA. 14-
DIRECTION.AND.PLACEMENT 15-ACTOR 16-ACTOR.PERSON 17-ACTOR.NUMBER 18-X 19-ACTOR.ANIMACY 20-
ACTOR.COREFERENCE 21-UNDERGOER 22-UNDERGOER.PERSON 23-UNDERGOER.NUMBER 24-X.1 25-
UNDERGOER.ANIMACY 26-UNDERGOER.COREFERENCE 27-Translation.value 28-Region 29-
Participant.Number 30-Gender 31-Age 32-Ethnicity 33-Language.background)
1:
```

Step 9: Now you're ready to do your first Rbrul run

```
ONE-LEVEL ANALYSIS OF RESPONSE UNDERGOER.DATA.SAYS WITH PREDICTOR(S): FILE.NAME [random] and VERB [random] and CA. (4.01e-08) + UNDERGOER.COREFERENCE (9.45e-05) +
UNDERGOER.ANIMACY (0.00195) + UNDERGOER.PERSON.AND.NUMBER.COMBINED (0.0139)

$CA.
      factor logodds tokens Modified/Modified+Unmodified centered factor weight
          CA   0.367    429                        0.783                 0.591
 EYEGAZEONLY   0.288    256                        0.770                 0.572
        NOCA  -0.655    340                        0.568                 0.342

$UNDERGOER.PERSON.AND.NUMBER.COMBINED
   factor logodds tokens Modified/Modified+Unmodified centered factor weight
   SECOND   0.424     57                       0.807                 0.604
    FIRST   0.080    241                       0.809                  0.52
 SINGULAR  -0.504    727                       0.667                 0.377

$UNDERGOER.ANIMACY
           factor logodds tokens Modified/Modified+Unmodified centered factor weight
          ANIMATE   0.323    689                        0.762                  0.58
    OTHERANIMATES   0.189    122                        0.705                 0.547
 NONLOC-INANIMATE  -0.512    214                        0.537                 0.375

$UNDERGOER.COREFERENCE
         factor logodds tokens Modified/Modified+Unmodified centered factor weight
 COREF(NULLS/O)   0.371    223                       0.834                 0.592
       COREF(N)   0.315     68                       0.750                 0.578
     COREF(PRO)  -0.152    112                       0.777                 0.462
  NOCOREFERENCE  -0.533    622                       0.646                  0.37
```