Multivariate ANOVA & Repeated Measures

Hanneke Loerts

April 16, 2008

Outline

- Introduction
- Multivariate ANOVA (MANOVA)
- Repeated Measures ANOVA
- Some data and analyses

Introduction

- When comparing two groups
  → T-test

- When comparing three or more groups
  → ANOVA

MANOVA

- Multivariate Analysis of Variance
  - Compares 3 or more groups
  - Compares variation between groups with variation within groups

- Difference: MANOVA is used when we have 2 or more dependent variables

An example

- Test effect of a new antidepressant (=IV)
  - Half of patients get the real drug
  - Half of patients get a placebo
- Effect is tested with BDI (=DV)
  - Beck Depression Index scores (a self-rated depression inventory)
- In this case → T-test

An example

- We add an independent variable
  - IV1 = drug type (drug or placebo)
  - IV2 = psychotherapy type (clinic or cognitive)
- We compare 4 groups now:
  - 1: placebo, cognitive
  - 2: drug, clinic
  - 3: placebo, clinic
  - 4: drug, cognitive
- In this case → ANOVA

An example

- We add two other dependent measures:
  - Beck Depression Index scores (a self-rated depression inventory),
  - Hamilton Rating Scale scores (a clinician rated depression inventory), and
  - Symptom Checklist for Relatives (a made up rating scale that a relative completes on the patient).

An example: the data

Group	Drug	Therapy	Mean BDI	Mean HRS	Mean SCR
1	Placebo	Cogn.	12	9	6
2	Drug	Clinic	10	13	7
3	Placebo	Clinic	16	12	4
4	Drug	Cogn.	8	3	2

Note: high scores indicate more depression, low scores indicate normality

Why not three separate ANOVA's?

- Increase in alpha-level → type 1 errors
- Univariate ANOVA's cannot compare the dependent measures

  → possible correlations are thrown away
- Use MANOVA

Recall:

- F statistic = MS_M / MS_R

- F statistic = total amount of variation that needs to be explained by:
  - MS_M = systematic variation / variance given that all observations come from single distribution
  - MS_R = residual variation / variance of each condition separately

Recall:

- F statistic = MS_M / MS_R

- If $F < 1 \rightarrow MS_R > MS_M$
- If $F > 1 \rightarrow MS_R < MS_M$

MANOVA

- Univariate ANOVA for every Dependent Variable
- But: we also want to know about the correlations between the DV's

MANOVA

- Each subject now has multiple scores: there is a matrix of responses in each cell
- Additional calculations are needed for the difference scores between the DV's
- Matrices of difference scores are calculated and the matrix squared
- When the squared differences are summed you get a sum-of-squares-and-cross-products-matrix
  - This is actually the matrix counterpart to the sums of squares
- Now we can test hypotheses about the effects of the IVs on linear combination(s) of the DVs

MANOVA

- Tests used for MANOVA:
  - Pillai's
  - Wilks'
  - Hotelling's

Hypotheses MANOVA

- H_0: There is no difference between the levels of a factor
- H_a: There is a difference between at least one level and the others

Assumptions MANOVA

- Independence of observations
- Multivariate normality
  - For dependent variables
  - For linear combinations
- Equality of covariance matrices (similar to homogeneity of variance)

Back to the example

- The effect of drug (IV1) and psychotherapy (IV2) on depression measures

- Now we add measurement points
  - Before the treatment
  - 1 week after the treatment
  - 2 weeks after the treatment
  - Etc.

Repeated measures

- When the same variable is measured more than once for each subject

- Reduces unsystematic variability in the design → greater power to detect effects

Repeated measures

- Violates the independence assumption
  - One subject is measured repeatedly

- Assumption of sphericity
  - relationship between pairs of experimental conditions is similar → level of dependence is roughly equal

Repeated measures

- Sphericity assumption ε
- Holds when:

  variance A-B = variance A-C = variance B-C

- Measured by Mauchly's test in SPSS
- If significant then there are differences and sphericity assumption is not met

MANOVA vs Repeated Measures

- In both cases: sample members are measured on several occasions, or trials
- The difference is that in the repeated measures design, each trial represents the measurement of the same characteristic under a different condition

MANOVA vs Repeated measures

- MANOVA: we use several dependent measures
  - BDI, HRS, SCR scores
- Repeated measures: might also be several dependent measures, but each DV is measured repeatedly
  - BDI before treatment, 1 week after, 2 weeks after, etc.

An experiment using Repeated Measures

- ERP: event-related brain potentials
  - Changes of voltage in the brain that can be time-locked to a specific (linguistic) stimulus

- ERP:
  - Provides a timeline of processing
  - Can tell us at which point certain aspects of language are processed in the brain

Compare: correct to incorrect

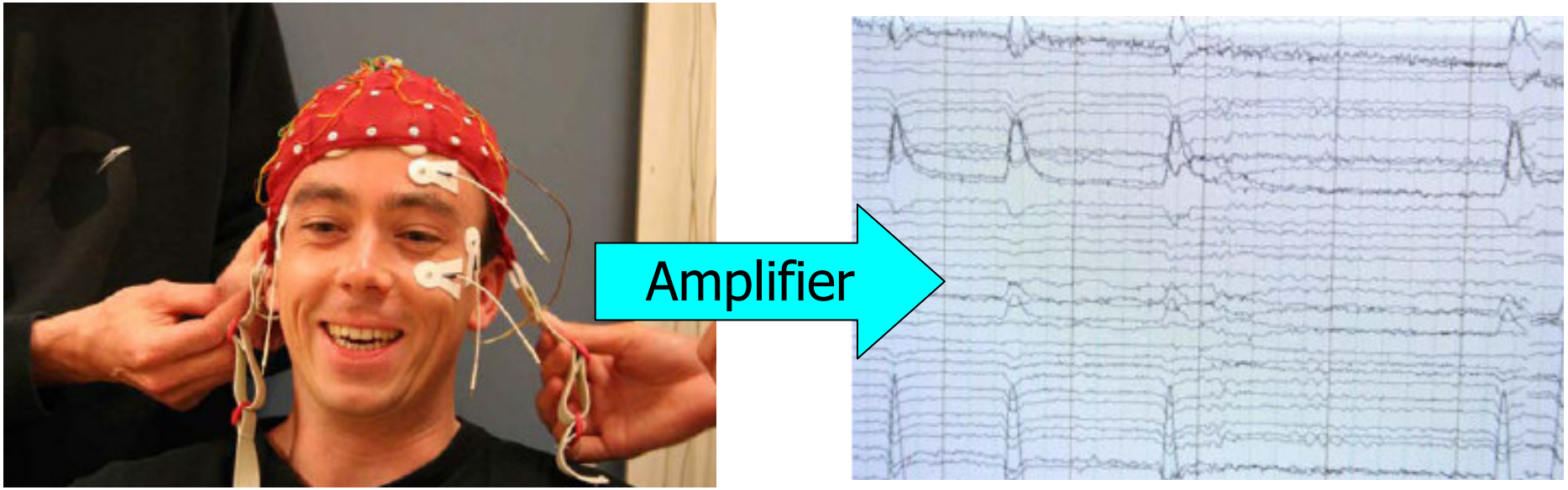
Event A: I saw a nice **cloud** on the horizon

Event B: I saw two nice **cloud** on the horizon

Compare: correct to incorrect

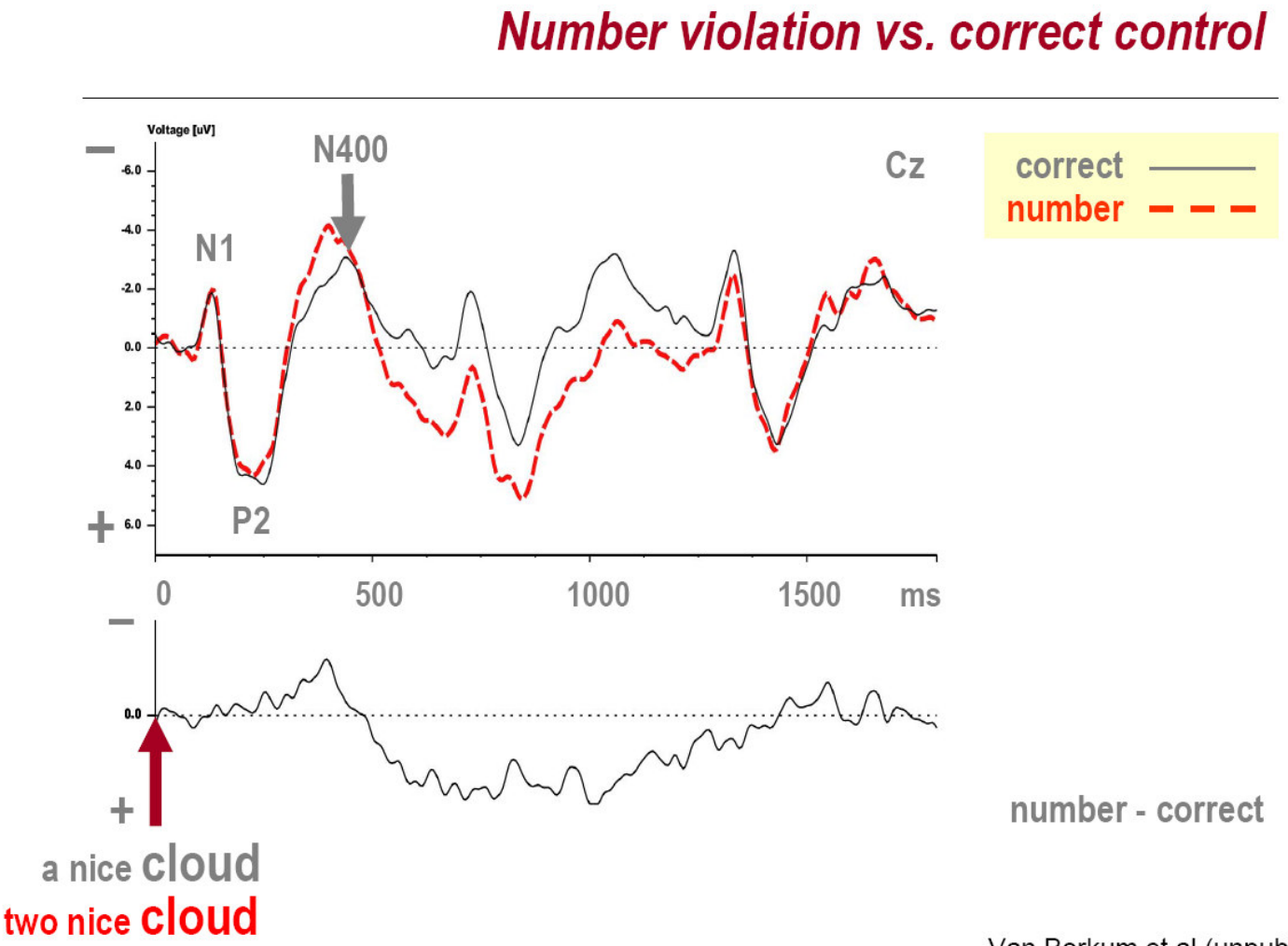
Event A: I saw a nice **cloud** on the horizon

Event B: I saw two nice **cloud** on the horizon



- Average EEG segments
  - For all subjects
  - For all event types

Result: ERP waveform associated with type A and type B



Van Berkum et al (unpublished)

What does this mean?

- Basic assumption: difficult condition elicits more activation
- Difference between two conditions reveals when the particular aspect (violation) is processed

This experiment

- Effect of word frequency
  - High versus low
- Effect of grammaticality
  - Grammatical versus ungrammatical
- 2 x 2 design

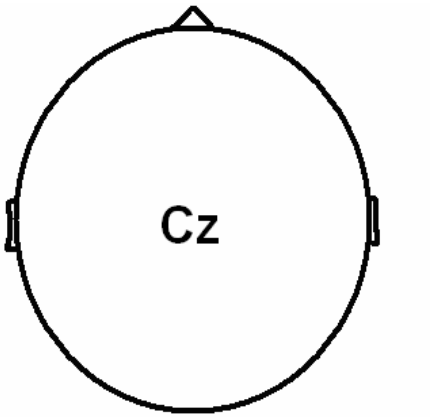
Background: Frequency

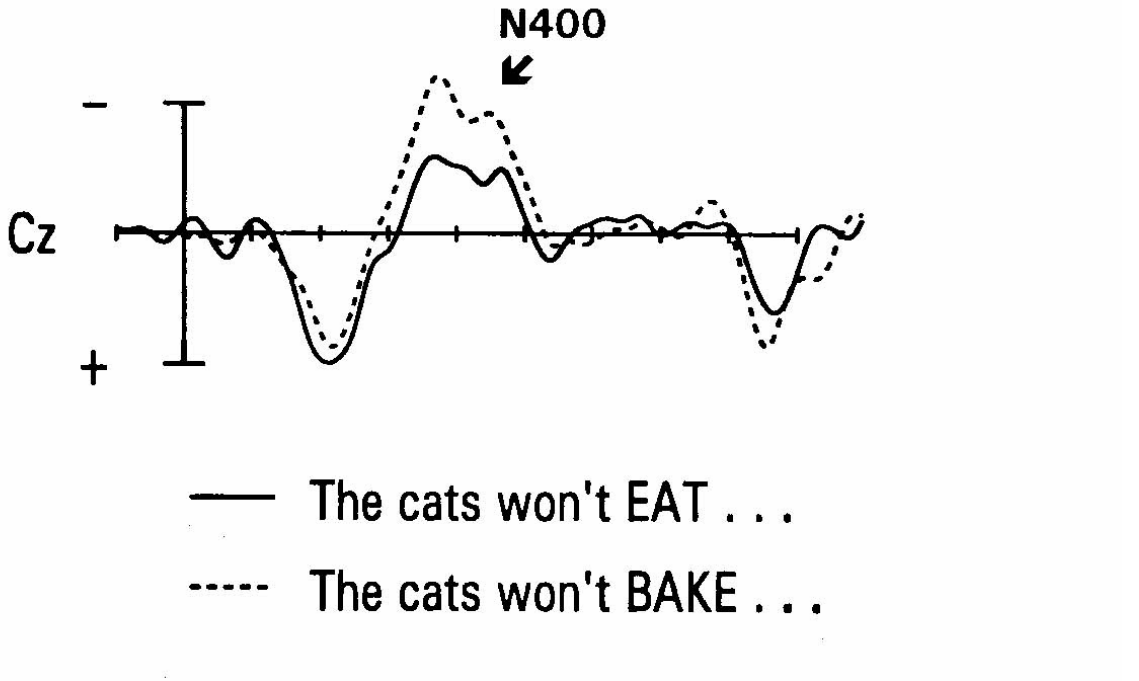
- Behavioural:
  - RT: faster to high frequency words
  - Frequency facilitates processing

- ERP:
  - Negative peak at 400 ms for low frequency
  - Low frequency words are more difficult

N400 frequency effect

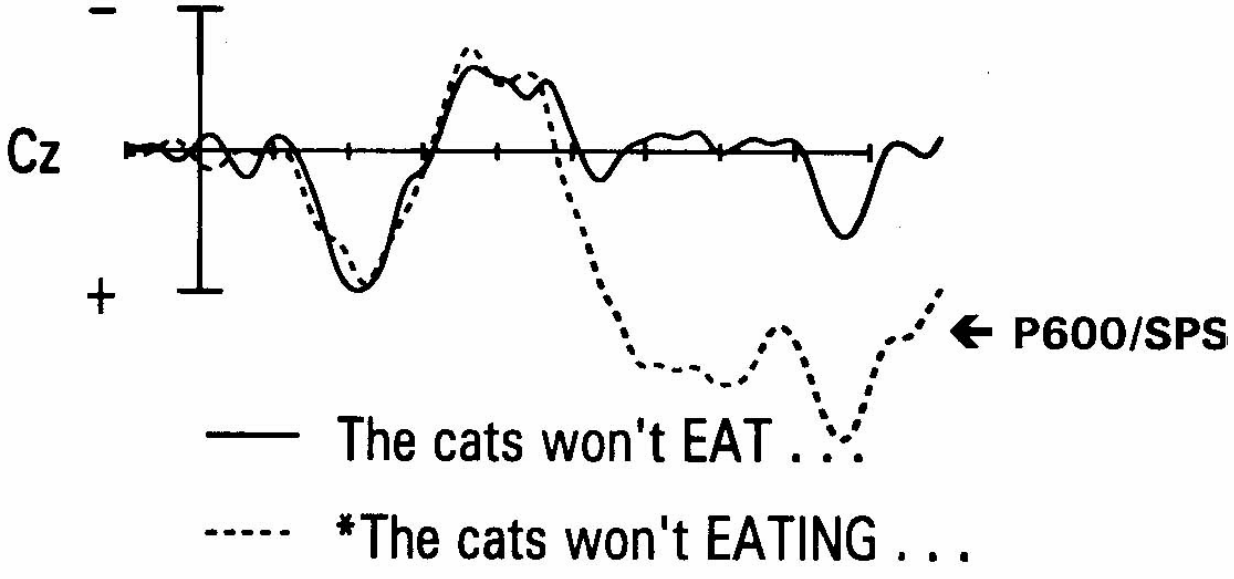
- Negativity for LF at 400 ms
- Related to semantic aspects
- Integration difficulty



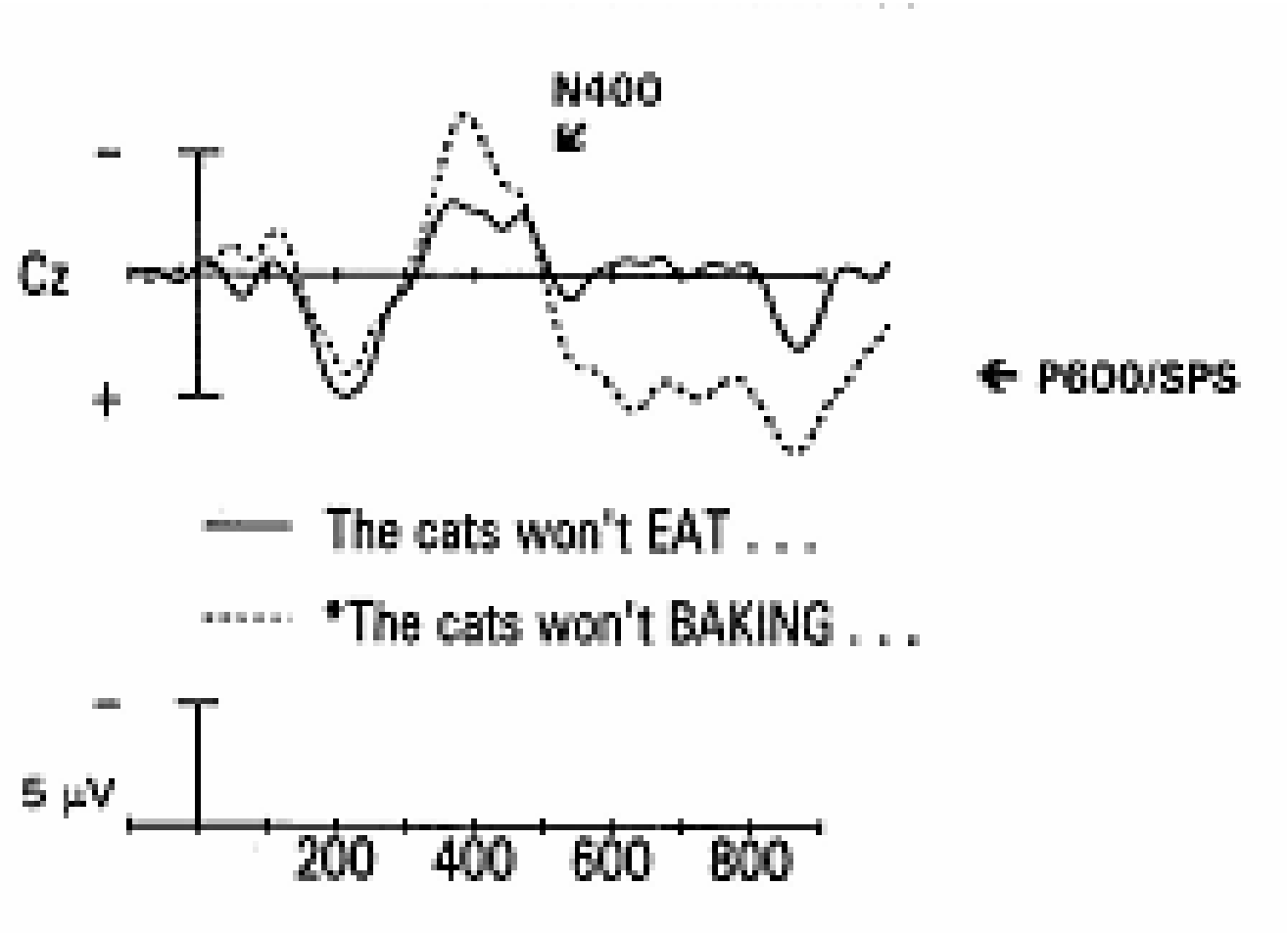


Processing syntax

- Detection of violation: early negativity
  - Left frontal
  - 300 ms
- Repair/re-analysis of violation: late positivity
  - Posterior
  - 600 ms



Semantics - Syntax



Present study

- ERP: time-line and stages of processing
- Violations of subject-verb agreement
  - *‘*he mow the lawn’*
  - Detection point around 300 ms
  - P600 for repair/re-analysis
- Additional factor: lexical frequency
  - E.g. *‘work’* vs *‘sway’*
  - N400 for low frequency
- Interaction?

Methods

- 160 experimental sentences

Freq.	**Gramm.**	**Example**
High	Correct	The scientist does not understand the new scales and **he calls** his wife for help.
	Incorrect	The scientist does not understand the new scales and ***he call** his wife for help.
Low	Correct	Marnix fell with his nose on the table and **he halts** the nose bleed with a tissue.
	incorrect	Marnix fell with his nose on the table and ***he halt** the nose bleed with a tissue.

Methods

- Matched on plausibility
- Matched on complexity
- Matched on frequency of surrounding words
- Matched on length of surrounding words
- Different lists
- Fillers: 224
- Questions in between

Methods

- 30 subjects
  - Age 18-26
  - Native Dutch
  - Right-handed
  - No neurological complaints
- In front of a screen
- Word by word presentation

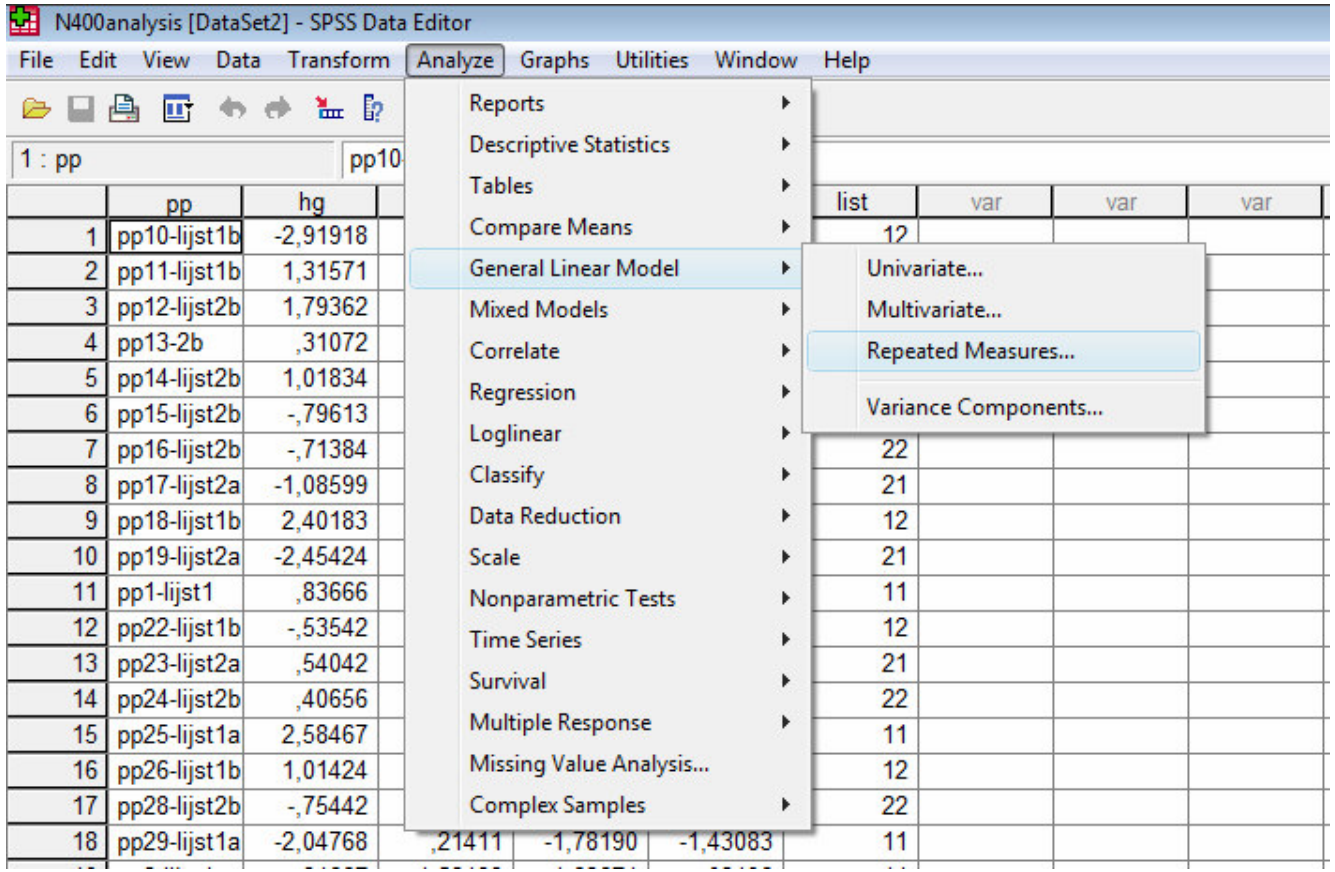
Hypotheses

- Low frequency verbs will be more difficult to process compared to high frequency verbs → N400
- Ungrammatical verbs will elicit a repair/reanalysis process → P600
- High frequency ungrammatical verbs might be detected with greater ease than low frequency ungrammatical verbs (around 300 ms → LAN)

Statistical analysis

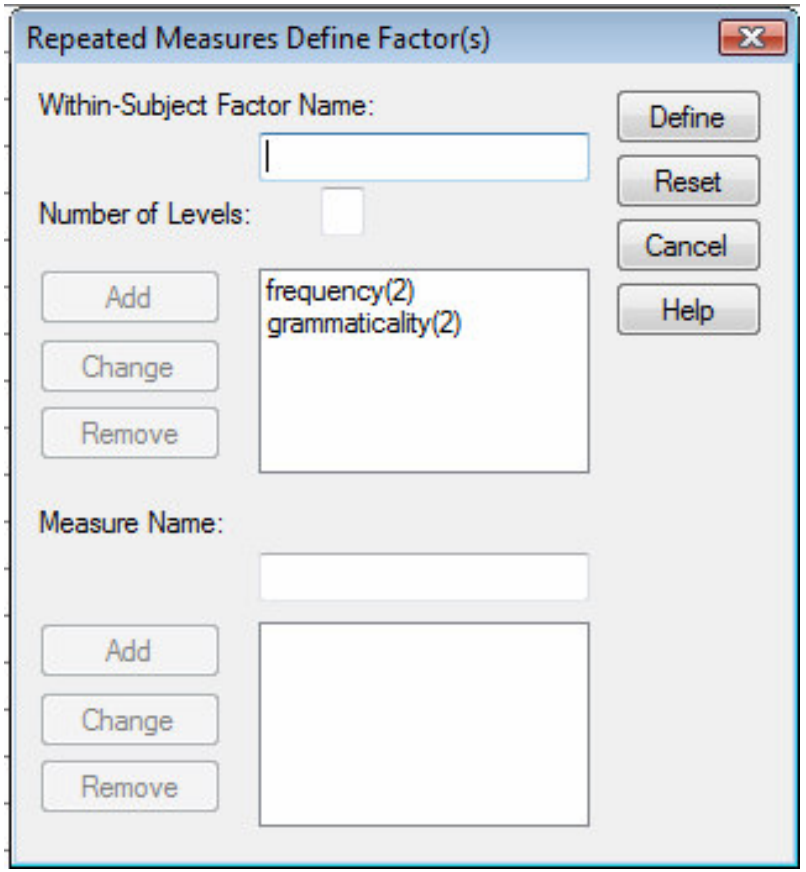
- Repeated measures ANOVA
  - Subjects are confronted with both grammaticality and frequency repeatedly
- Test equality of means
- Mean raw amplitude scores in SPSS

Data analysis



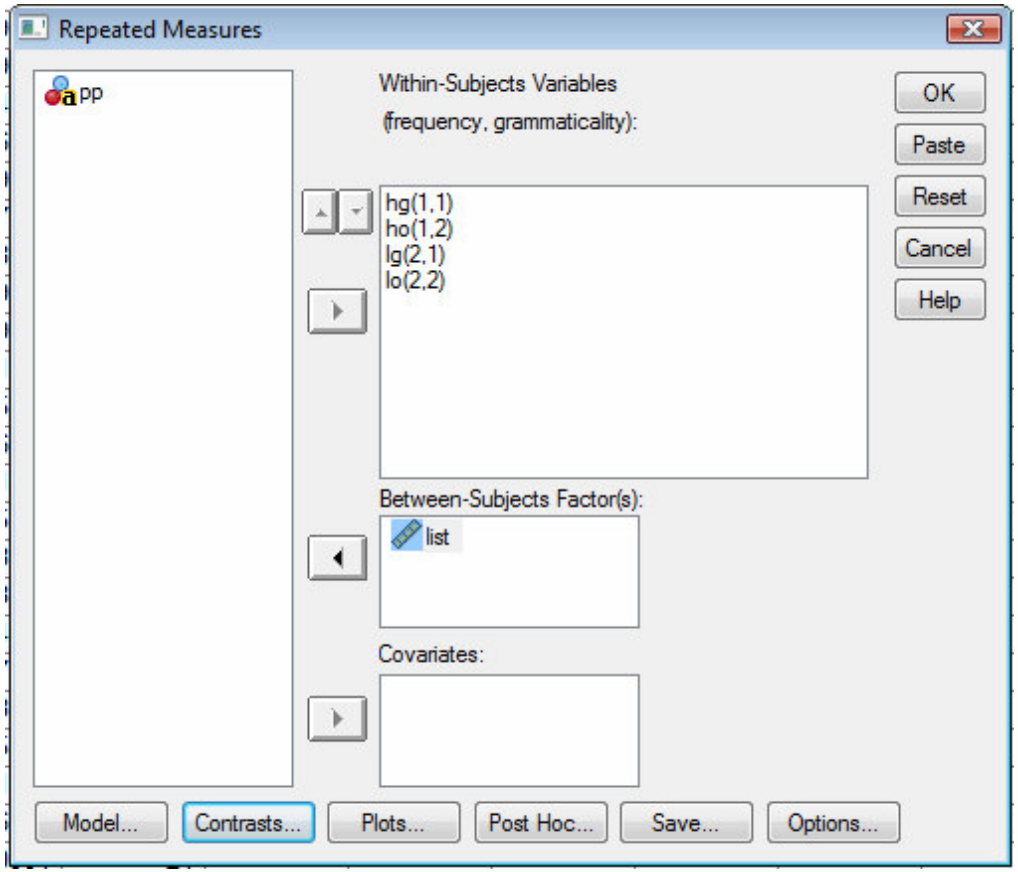
Data analysis

- Repeated measures or Within-Subject Factors:
  - Frequency (2)
  - Grammaticality (2)



Data analysis

Between-Subjects Factor: List



What we expected:

- Frequency effect → N400
- Grammaticality effect → P600
- Difference in detection → interaction

Results: N400

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		35,968	1	35,968	21,006	,000
frequency * list	Linear		1,472	3	,491	,287	,835
Error(frequency)	Linear		44,518	26	1,712		
gramm		Linear	,184	1	,184	,135	,716
gramm * list		Linear	1,856	3	,619	,455	,716
Error(gramm)		Linear	35,333	26	1,359		
frequency * gramm	Linear	Linear	4,593	1	4,593	3,095	,090
frequency * gramm * list	Linear	Linear	6,793	3	2,264	1,526	,231
Error(frequency*gramm)	Linear	Linear	38,580	26	1,484		

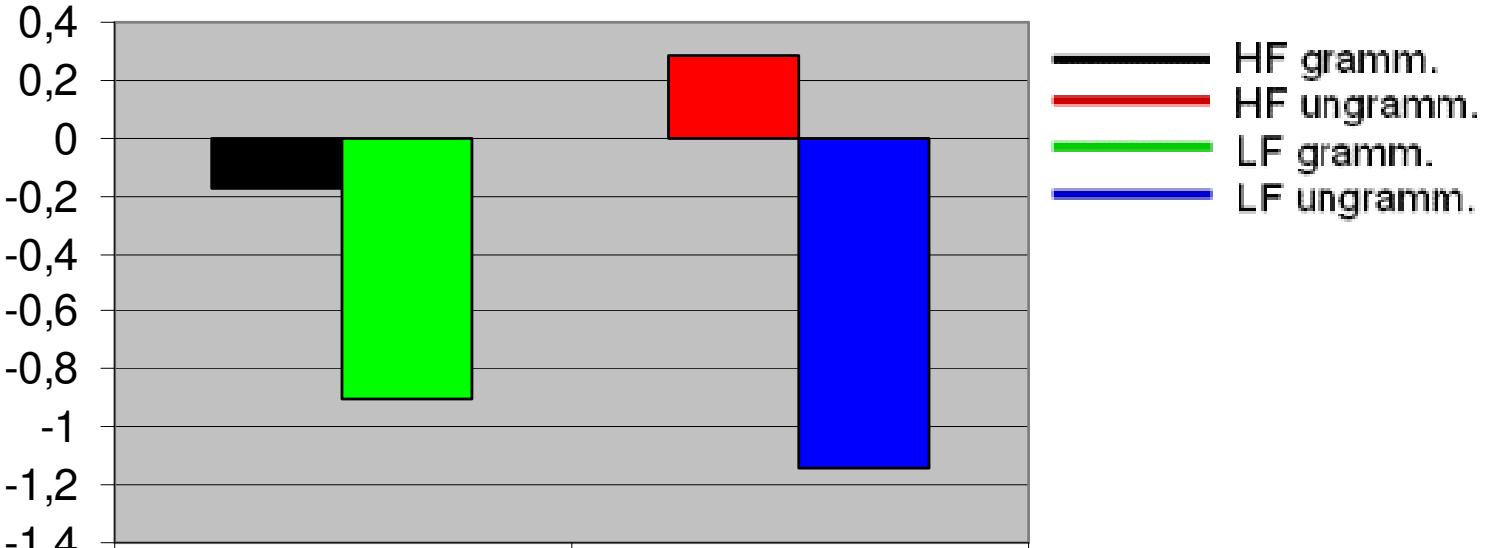
Results: N400

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		35,968	1	35,968	21,006	,000
frequency * list	Linear		1,472	3	,491	,287	,835
Error(frequency)	Linear		44,518	26	1,712		
gramm		Linear	,184	1	,184	,135	,716
gramm * list		Linear	1,856	3	,619	,455	,716
Error(gramm)		Linear	35,333	26	1,359		
frequency * gramm	Linear	Linear	4,593	1	4,593	3,095	,090
frequency * gramm * list	Linear	Linear	6,793	3	2,264	1,526	,231
Error(frequency*gramm)	Linear	Linear	38,580	26	1,484		

Results: N400



	gr	ungr
HF	-0,178658889	0,287041111
LF	-0,905774074	-1,140402222

Results: P600

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		,117	1	,117	,066	,800
frequency * list	Linear		3,314	3	1,105	,621	,608
Error(frequency)	Linear		46,273	26	1,780		
gramm		Linear	68,725	1	68,725	33,832	,000
gramm * list		Linear	2,138	3	,713	,351	,789
Error(gramm)		Linear	52,815	26	2,031		
frequency * gramm	Linear	Linear	5,924	1	5,924	6,321	,018
frequency * gramm * list	Linear	Linear	5,826	3	1,942	2,072	,128
Error(frequency*gramm)	Linear	Linear	24,367	26	,937		

Results: P600

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		,117	1	,117	,066	,800
frequency * list	Linear		3,314	3	1,105	,621	,608
Error(frequency)	Linear		46,273	26	1,780		
gramm		Linear	68,725	1	68,725	33,832	,000
gramm * list		Linear	2,138	3	,713	,351	,789
Error(gramm)		Linear	52,815	26	2,031		
frequency * gramm	Linear	Linear	5,924	1	5,924	6,321	,018
frequency * gramm * list	Linear	Linear	5,826	3	1,942	2,072	,128
Error(frequency*gramm)	Linear	Linear	24,367	26	,937		

Interaction?

- The end-effect of the N400?
- Split up the time-windows:
  - 450-600 for the onset
  - 600-1000 for the ‘real’ P600

- Look at the effects separately

The 450-600 time-window

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		14,898	1	14,898	9,211	,005
frequency * list	Linear		2,612	3	,871	,538	,660
Error(frequency)	Linear		42,052	26	1,617		
gramm		Linear	5,863	1	5,863	3,407	,076
gramm * list		Linear	3,079	3	1,026	,596	,623
Error(gramm)		Linear	44,736	26	1,721		
frequency * gramm	Linear	Linear	10,228	1	10,228	6,706	,016
frequency * gramm * list	Linear	Linear	5,641	3	1,880	1,233	,318
Error(frequency*gramm)	Linear	Linear	39,656	26	1,525		

The 600-1000 time-window

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	frequency	gramm	Type III Sum of Squares	df	Mean Square	F	Sig.
frequency	Linear		,117	1	,117	,066	,800
frequency * list	Linear		3,314	3	1,105	,621	,608
Error(frequency)	Linear		46,273	26	1,780		
gramm		Linear	68,725	1	68,725	33,832	,000
gramm * list		Linear	2,138	3	,713	,351	,789
Error(gramm)		Linear	52,815	26	2,031		
frequency * gramm	Linear	Linear	5,924	1	5,924	6,321	,018
frequency * gramm * list	Linear	Linear	5,826	3	1,942	2,072	,128
Error(frequency*gramm)	Linear	Linear	24,367	26	,937		

What does the interaction mean?

- We expected a difference in the detection around 300 ms
- Instead there seems to be a difference in the onset of the P600 (based on raw data)
- To find out what the onset difference is → separate ANOVA's for high and low frequency verbs

What does the interaction mean?

When only taking high frequency verbs: grammaticality effect

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
gramm	Sphericity Assumed	15,790	1	15,790	6,656	,016
	Greenhouse-Geisser	15,790	1,000	15,790	6,656	,016
	Huynh-Feldt	15,790	1,000	15,790	6,656	,016
	Lower-bound	15,790	1,000	15,790	6,656	,016
gramm * list	Sphericity Assumed	,245	3	,082	,034	,991
	Greenhouse-Geisser	,245	3,000	,082	,034	,991
	Huynh-Feldt	,245	3,000	,082	,034	,991
	Lower-bound	,245	3,000	,082	,034	,991
Error(gramm)	Sphericity Assumed	61,675	26	2,372		
	Greenhouse-Geisser	61,675	26,000	2,372		
	Huynh-Feldt	61,675	26,000	2,372		
	Lower-bound	61,675	26,000	2,372		

What does the interaction mean?

When only taking high frequency verbs: grammaticality effect

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
gramm	Sphericity Assumed	15,790	1	15,790	6,656	,016
	Greenhouse-Geisser	15,790	1,000	15,790	6,656	,016
	Huynh-Feldt	15,790	1,000	15,790	6,656	,016
	Lower-bound	15,790	1,000	15,790	6,656	,016
gramm * list	Sphericity Assumed	,245	3	,082	,034	,991
	Greenhouse-Geisser	,245	3,000	,082	,034	,991
	Huynh-Feldt	,245	3,000	,082	,034	,991
	Lower-bound	,245	3,000	,082	,034	,991
Error(gramm)	Sphericity Assumed	61,675	26	2,372		
	Greenhouse-Geisser	61,675	26,000	2,372		
	Huynh-Feldt	61,675	26,000	2,372		
	Lower-bound	61,675	26,000	2,372		

What does the interaction mean?

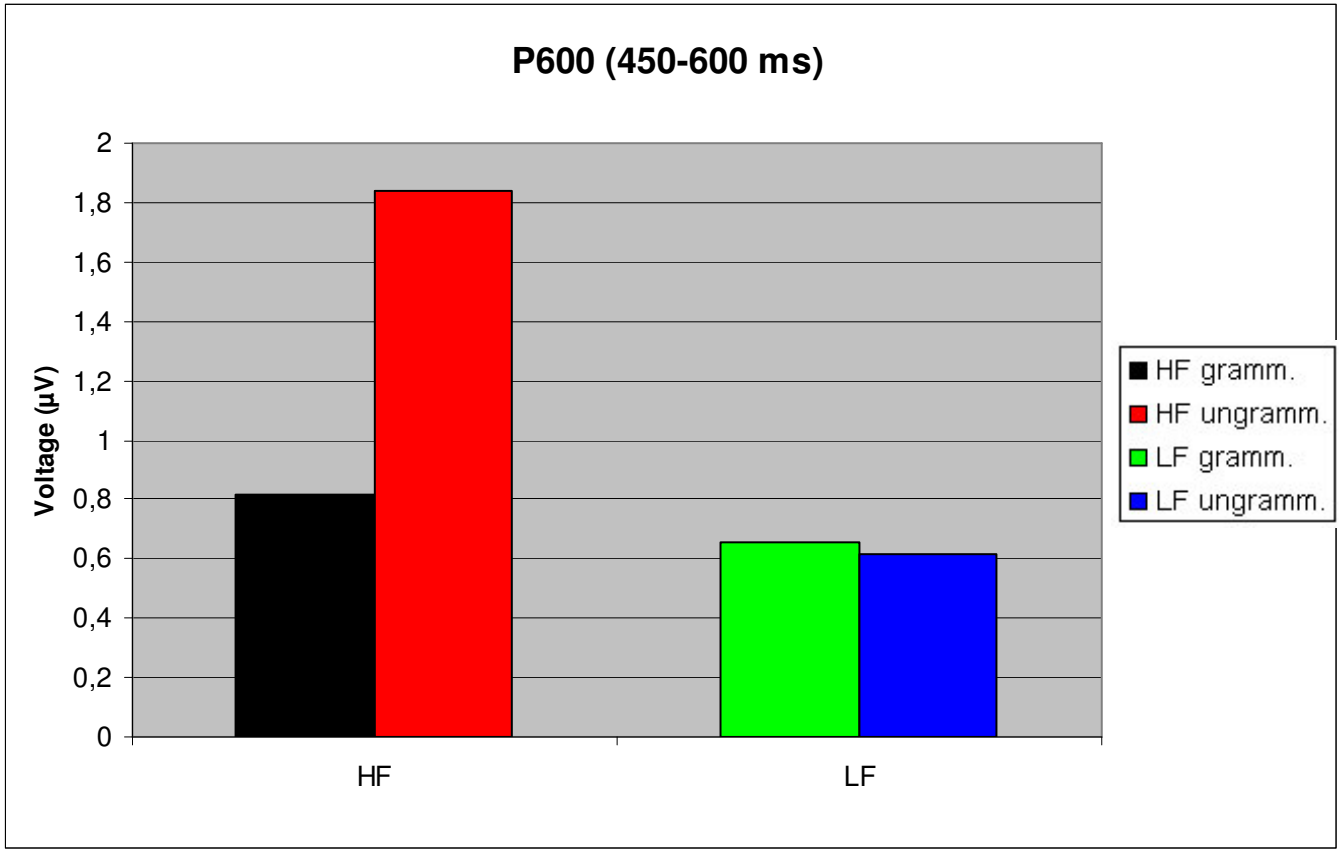
When only taking low frequency verbs: NO grammaticality effect

Tests of Within-Subjects Effects

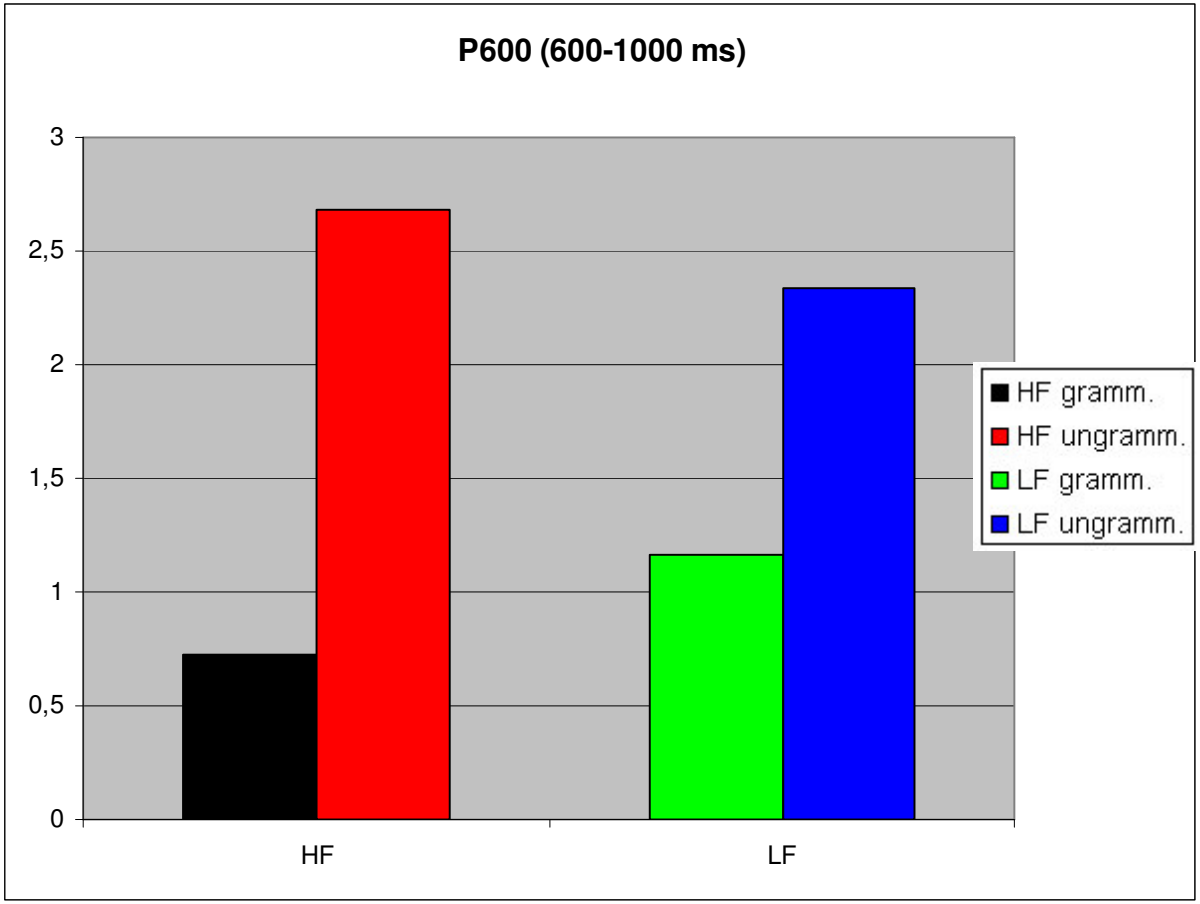
Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
gramm	Sphericity Assumed	,302	1	,302	,345	,562
	Greenhouse-Geisser	,302	1,000	,302	,345	,562
	Huynh-Feldt	,302	1,000	,302	,345	,562
	Lower-bound	,302	1,000	,302	,345	,562
gramm * list	Sphericity Assumed	8,474	3	2,825	3,233	,039
	Greenhouse-Geisser	8,474	3,000	2,825	3,233	,039
	Huynh-Feldt	8,474	3,000	2,825	3,233	,039
	Lower-bound	8,474	3,000	2,825	3,233	,039
Error(gramm)	Sphericity Assumed	22,717	26	,874		
	Greenhouse-Geisser	22,717	26,000	,874		
	Huynh-Feldt	22,717	26,000	,874		
	Lower-bound	22,717	26,000	,874		

The ‘real’ data



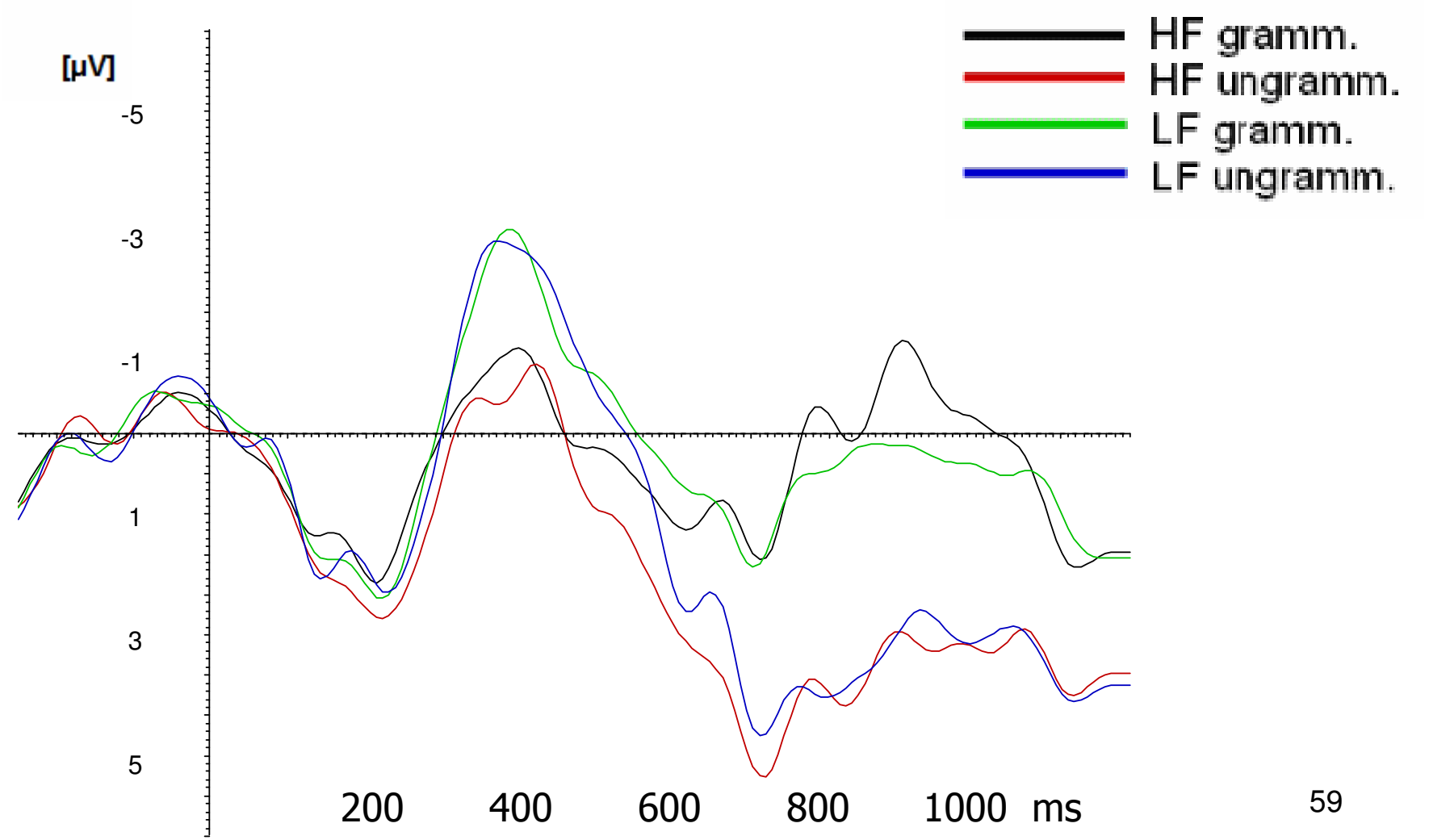
The ‘real’ data



Results

- When comparing high and low frequency
  - N400: negativity for low frequency
- When contrasting grammaticality
  - P600: positivity for ungrammatical
  - But: no early detection around 300 ms

Results


[µV]
-5
-3
-1
1
3
5
200
400
600
800
1000
ms
HF gramm.
HF ungramm.
LF gramm.
LF ungramm.

Discussion: Why no detection?

- Due to rules of different languages
  - '(…) he mows/*mow the lawn'
  - '(…) hij roept/*hij roep (he calls/*call)
  - Word order issue?
- Due to strictness of violated rule
  - 'The scientist criticized Max's **of** proof…'
  - More obvious: earlier detection?

Conclusion

- Frequency and grammaticality elicit different brain responses
- High frequency verbs are more easily processed than low frequency verbs
- People initialize a repair process after 600 ms when confronted with subject-verb agreement violations

Conclusion

- The repair process can be initialized earlier when the ungrammatical verb is a high frequency one compared to a low frequency