Eye tracking in linguistics, with a focus on the Visual World Paradigm from simple to hard

Björn Lundquist (UiT)

The eye tracker in linguistic research

- Visual World Paradigm: Looking at pictures while listening to stimuli that is related to the pictures in some way or other.
- Potentially very simple (RQ1): does the participant understand the spoken stimulus, as revealed by fixating the picture stimulus that best matches the spoken stimulus.
- But often very complicated (RQ2): "does speaker group X show similar fixation patterns as speaker group Y when presented with sound/image stimulus of type A, or only a subset of A. (e.g. Can L2 speakers make use of gender markers on a prenominal modifier to locate the target noun, and in such case, is there furthermore an effect of gender congruency?)

Test if a group of participant can link (some aspect of) a linguistic form to a specific meaning:

- Example 1: test if infants know a set of words.
- Show two pictures while repeating the name of one of then, e.g. "ball, ball, ball, ball", and in a subsequent trial, repeat the name of the other object.

- Analysis simple: compare the fixations to the target to the fixations to the competitor/non-target, measured in ms or actual fixations, over the whole trial.
- No need to time-lock a specific region of interest to the sound stimulus.

Test if a group of participant can link (some aspect of) a linguistic form to a specific meaning:

- Example 2: Do children understand case marking or voice marking?
- Picture stimulus: picture of transitive events with contrasting argument structure relations, e.g. *A cat chasing a dog*, and *a dog chasing a cat*.
- Sound stimulus.

(1) The dog was chased by the cat / The cat chased the dog.

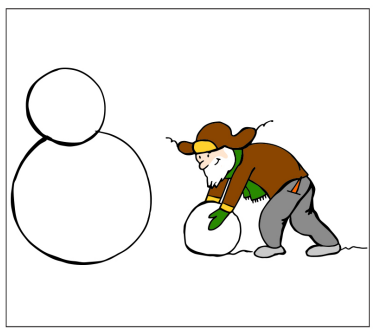
(2) The dog(ACC) chased the cat(NOM)/The dog(NOM) chased the cat(ACC)

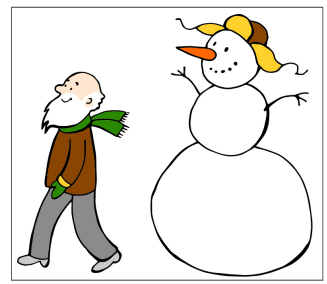
- Analysis simple: compare the fixations to the intended target to the fixations to the competitor/non-target, measured in ms or actual fixations, over the whole trial.
- No need to time-lock a specific region of interest to the sound stimulus.

Test if a group of participant can link (some aspect of) a linguistic form to a specific meaning:

- Example 3: Do adult speakers associate a specific meaning to a certain grammatical or lexical item?
- Do English speakers associate the simple past tense with "perfective"/complete event? (see Minor et al.)

(3) Grandpa build a snowman.





- Analysis potentially simple: compare the fixations to the intended target to the fixations to the competitor/non-target, measured in ms or actual fixations, over the whole trial, or from the offset of the relevant grammatical marker.

Why things get more complicated

- In the examples above, we don't presuppose that the participants will fixate the intended target, rather this is our RQ.
- The set-up and the following analysis are relatively simple: we can directly compare the number of/proportion of fixations to target vs competitor/non-target.
- Many VWP-studies, especially in the field of bilingualism and language acquisition, target **predictive processing/incremental parsing**.

Two classic studies

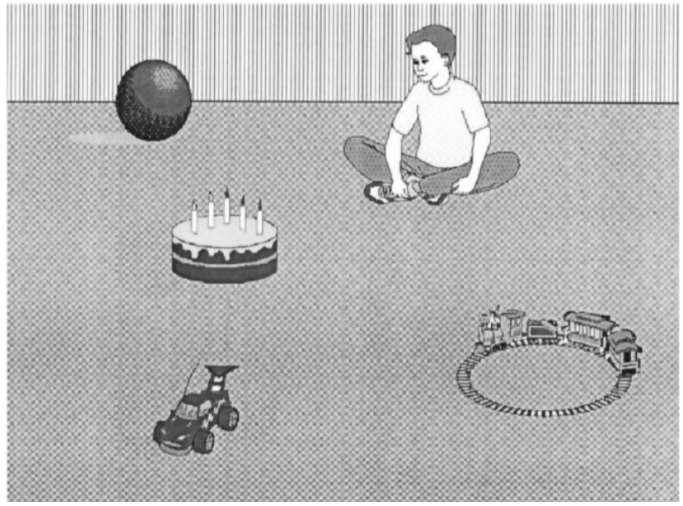
Grüter, Lew-Williams and Fernald 2013, gender in Spanish:

Encuentra...

la pelota? (el zapato)

Altmann and Kamide 1999, predictions based on subcategorization frames:

The boy will eat the cake



- The participants will surely fixate the target to a higher degree that the competitors in both these cases.
- But we are only interested in the the fixations that were triggered by the grammatical/lexical cue under investigation.

What's the problem

(4) encuentra la pelota
500ms 300ms 700ms
"Find the ball"

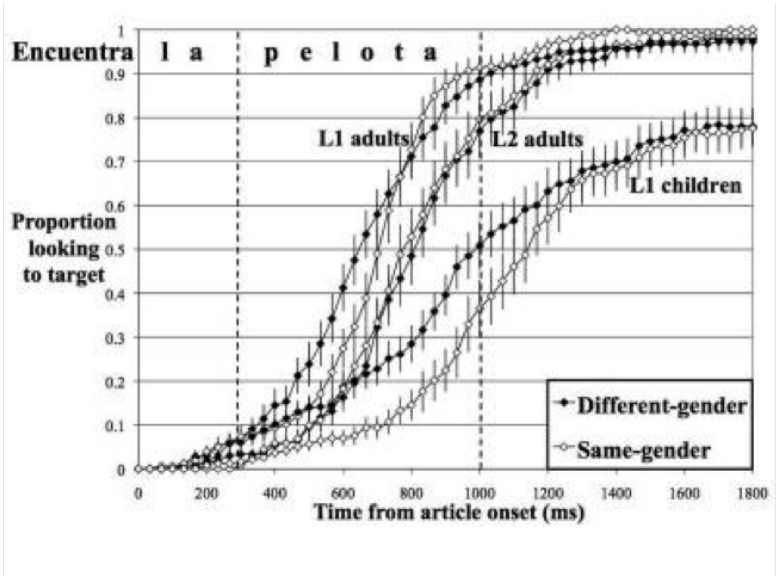
- There's a 300ms window from onset of gender article to onset of noun – could we focus on this time slice only, and compare looks to target and looks to competitor.
- Participants vary in general speed (processing, saccades, hearing). We don't really know what the right window is.
- Participants fixation patterns are influenced by a range of factors, of which the grammatical cue only is one.
- **A direct comparison of looks to target and competitor within a narrow time frame is not an ideal solution.**
- Alternative: compare the target fixation in a constraining context (experimental condition) and a non-constrained condition.

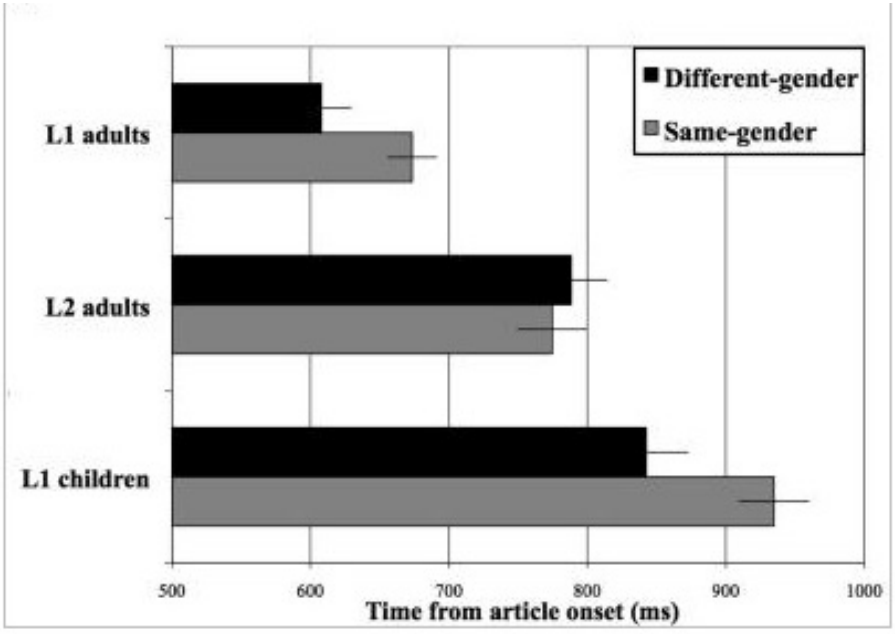
Grüter, Lew-Williams and Fernald 2013, L2 and Gender

Encuentra...

la pelota? (la galleta) Same

la pelota? (el zapato) Diff.

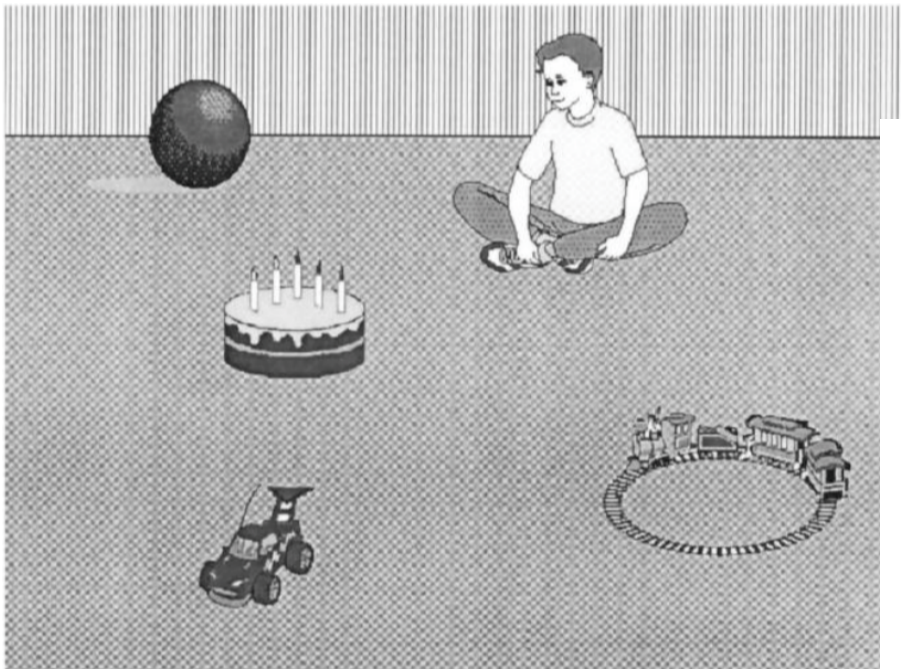


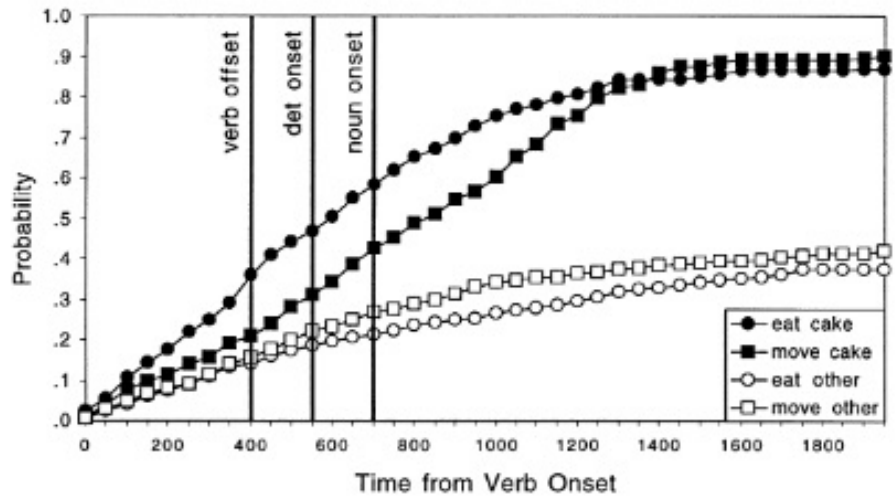


Altmann and Kamide 1999

(5) The boy will move the cake.

(6) The boy will eat the cake.

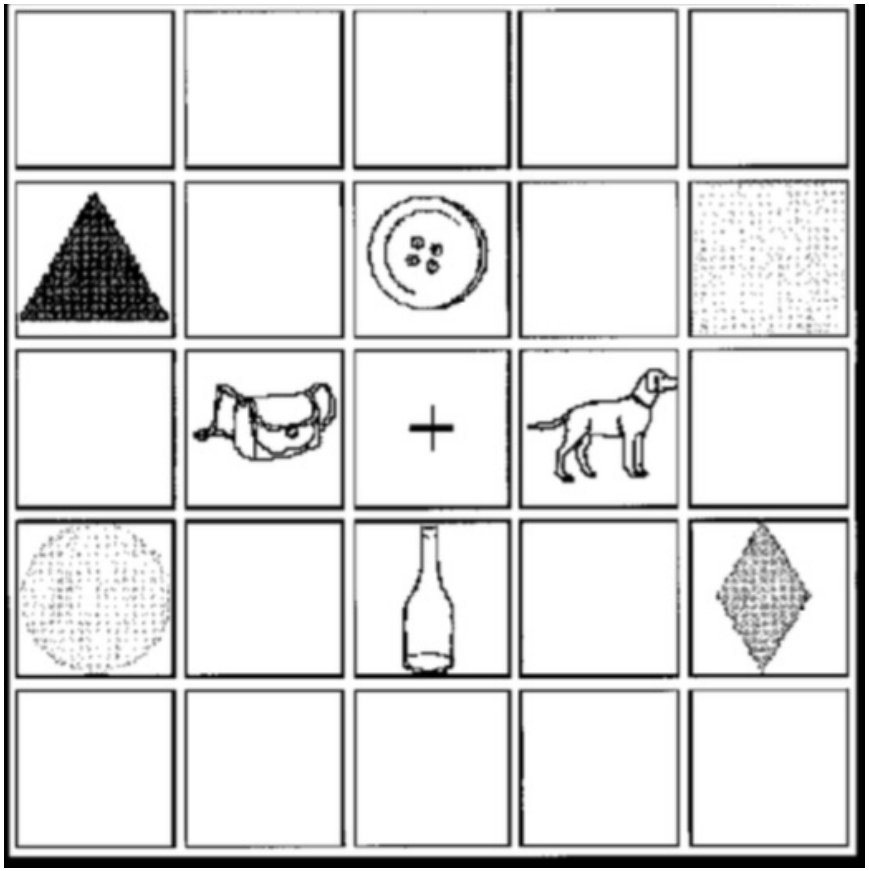




How to construe experimental and control condition

1. Change visual stimulus, but keep the same sound stimulus – standard in gender paradigms (Grüter, Lew-Williams and Fernald 2013, Hopp 2014 etc.).
2. Keep the visual stimulus, but change the sound stimulus, as in Altmann and Kamide 1999, but also in grammatical gender studies.

Dahan et al. 2000: The role of grammatical gender



French : No grammatical gender distinction on plural definite determiners:

Cliquez sur les boutons
Cliquez sur les bouteilles
Gender on singular determiners

Cliquez sur **le** bouton
Cliquez sur **la** bouteille

How to construe experimental and control condition

1. Change visual stimulus, but keep the same sound stimulus – standard in gender paradigms (Grüter, Lew-Williams and Fernald 2013, Hopp 2014 etc.).
2. Keep the visual stimulus, but change the sound stimulus, as in Altmann and Kamide 1999, but also in grammatical gender studies.

Encuentra...

la pelota? (el zapato)

Encuentra...

la pelota? (la galleta) Same

Potential problems with comparing same/diff conditions

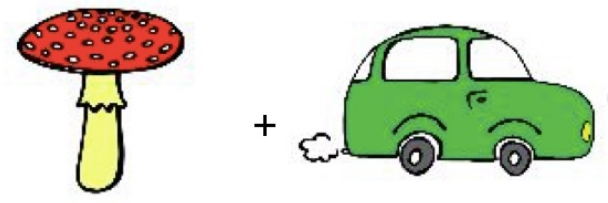
- Compared to directly comparing looks to target vs. looks to competitor, the more cautious way of comparing looks to target in condition A (constraining/experimental) and condition B (non-constraining/control), forces you to basically double the number of trials in the experiment.
- The effects may get artificially small: looks to target in the the control condition may still be guided by the grammatical cue (e.g. gender marking).

Temporal properties and effect sizes (data from Johannessen et al. subm.)

Jeg tenker på **en** *avbilda bil*

I'm thinking about a depicted car.

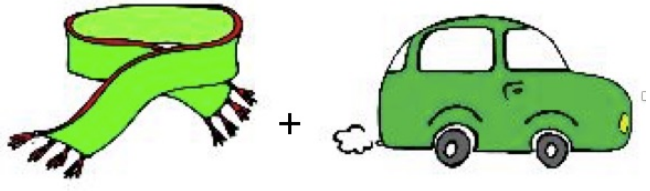
Same condition:



en sopp a.M mushroom

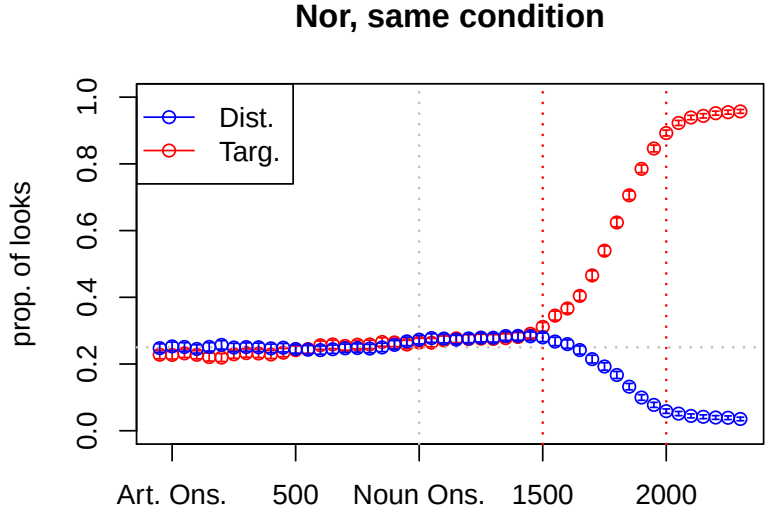
en bil (Targ.) a.M car

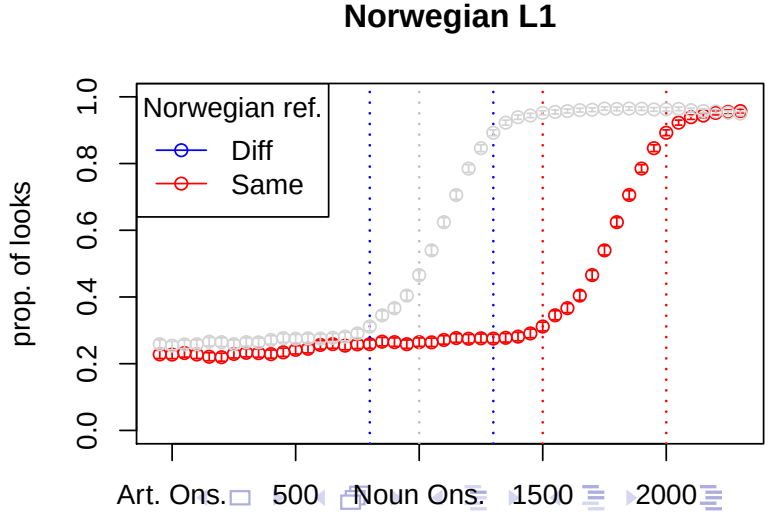
Different condition:



et skjerf a.N scarf

en bil (Targ.) a.M car

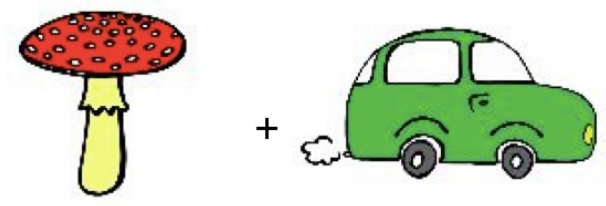




Johannessen et al. (in proc.): Gender processing in Norwegian L2 (and L1)

Jeg tenker på **en** *avbilda bil*

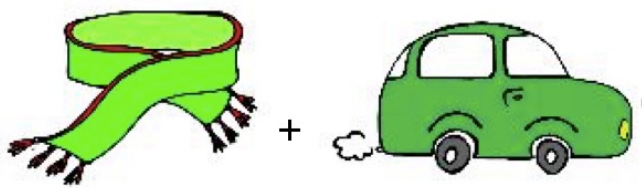
Same condition:



en sopp
a.M mushroom

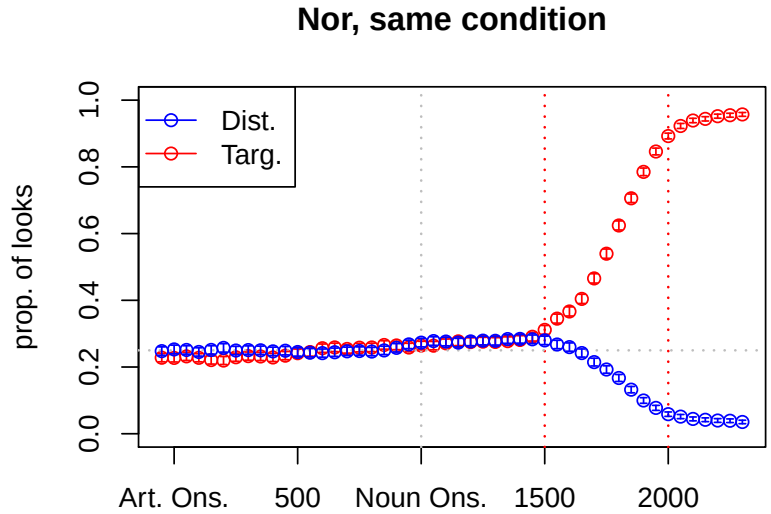
en bil (Targ.)
a.M car

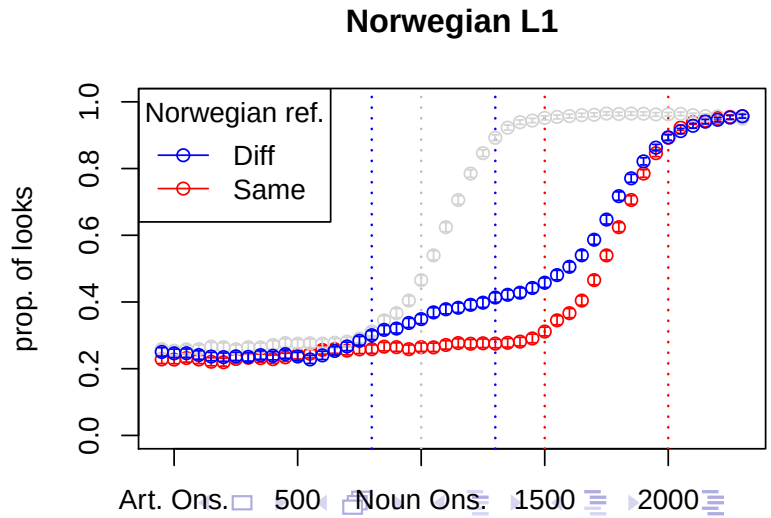
Different condition:



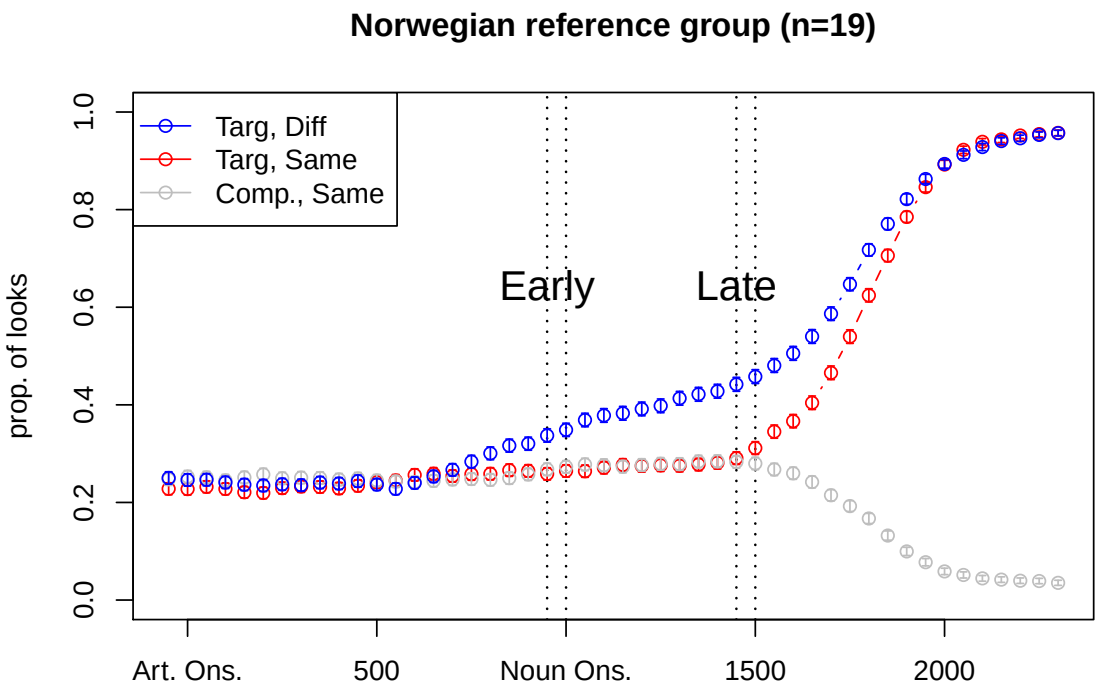
et skjerf
a.N scarf

en bil (Targ.)
a.M car

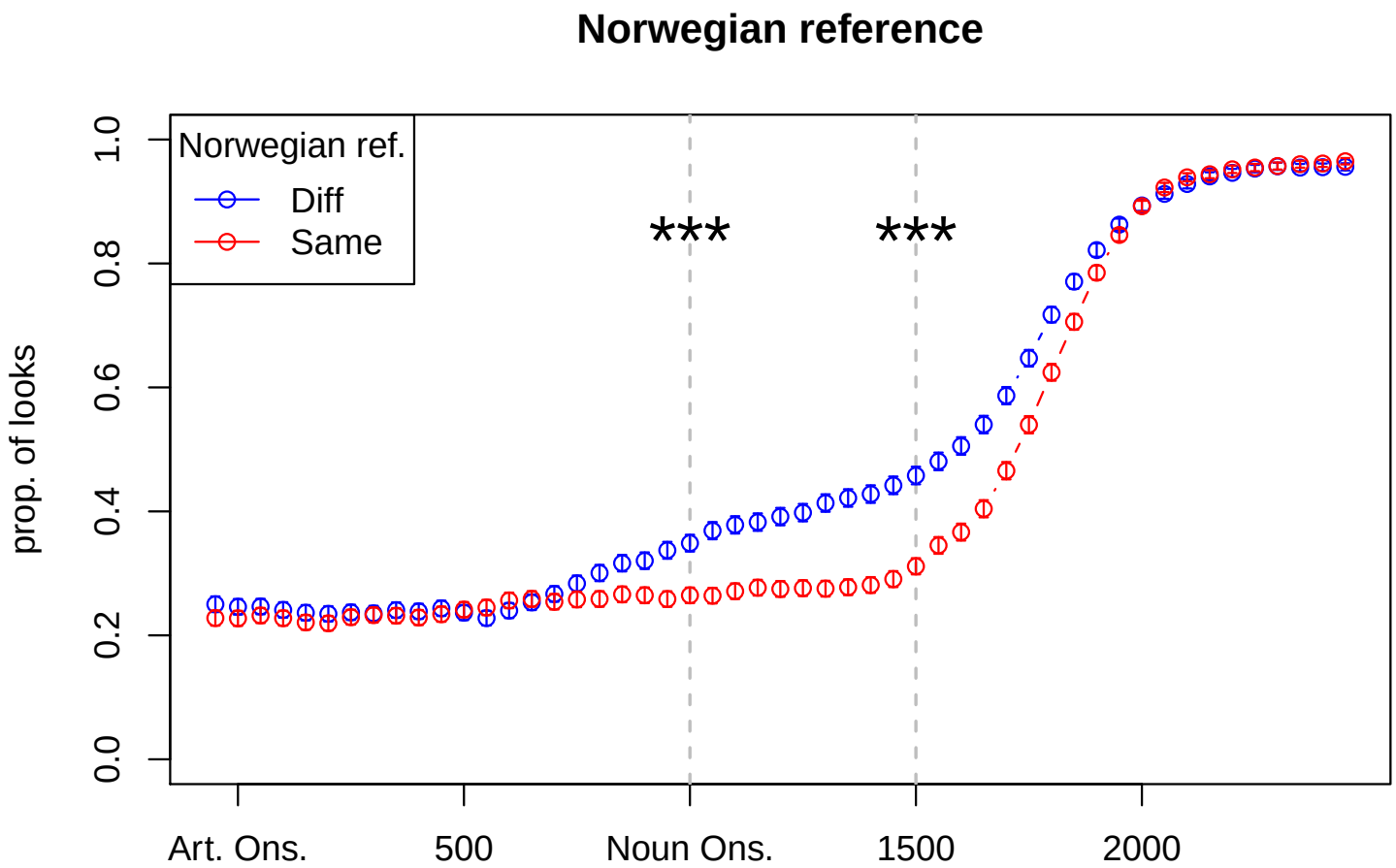




Establishing relevant ROI



A note on the statistical analysis



Glmer (binomial) at two 50ms slots. Dep. var: looks target (1/0).
glmer(look at target ~ condition + (1+cond I Speaker) + (1 + cond I Item))

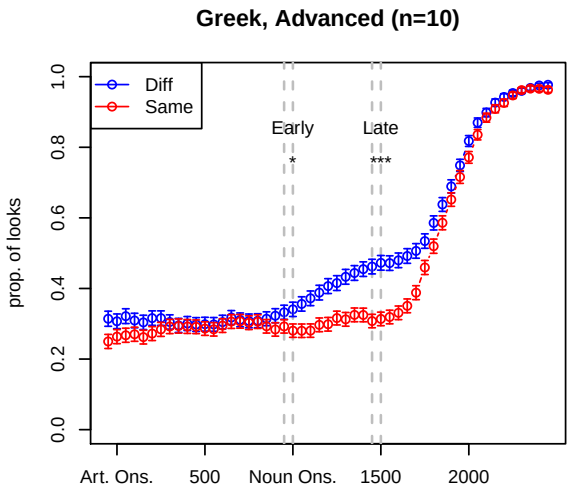
Comparing L1 to L2 processing – RQs

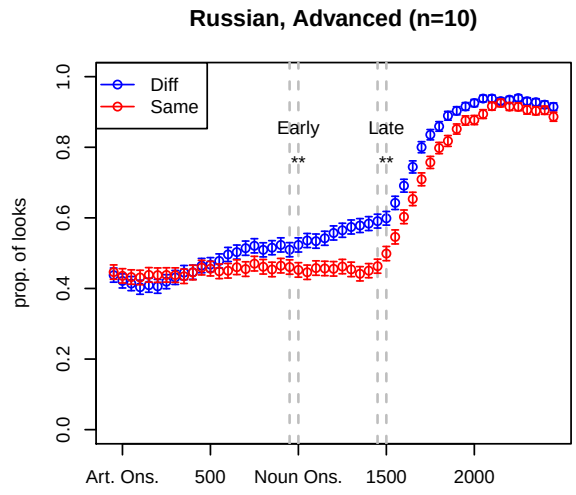
- CLI in processing: +/- presence of gender (Turkish vs. Greek/Russian), +/- presence of gender marked articles (Greek vs. Russian).
- The effect of general proficiency: Based on a naming task, the three participant groups were divided into an advanced and a high-intermediate group.

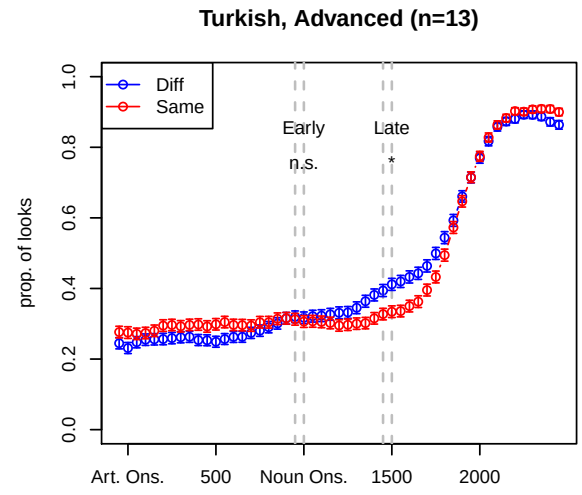
	L1 Greek	L1 Russian	L1 Turkish
Advanced	8	13	12
High-intermediate	15	10	8

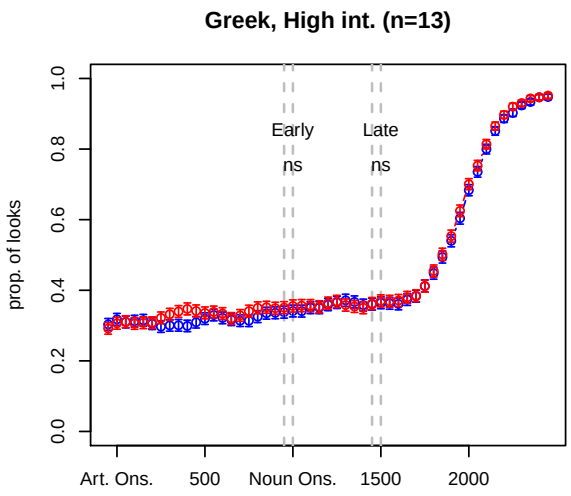
- Small groups but fully comparable to Dussias et al. 2014, Grüter et al, Hopp and Lemmerth etc.

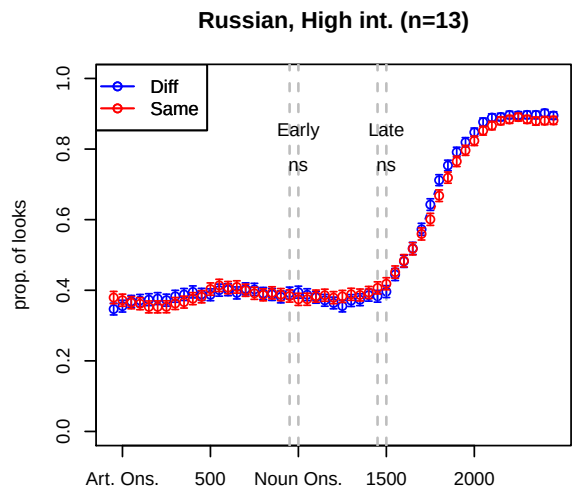
Comparing L2 results

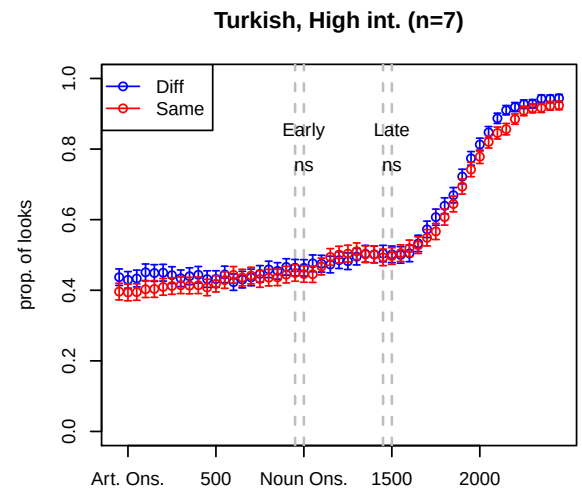












The complications, statistics

- How do make comparisons across groups?

glmer(look at target ~ condition * Lang * ProfGroup + (1+cond I Speaker) + (1 + cond I Item))

- Problems 1: L1 reference group only has one proficiency level.
- Problem 2: It's very hard to interpret the results from 3-ways interactions (but not impossible).
- Alternative: Just fit a model for the L2-groups, without reference to L1 group (solves the first problem, but to be hones, we do not find a three-way interaction.)
- Or code the L2 groups as six separate groups, which is what we did here: glmer(look at target ~ condition * Group + (1+cond I Speaker) + (1 + cond I Item))

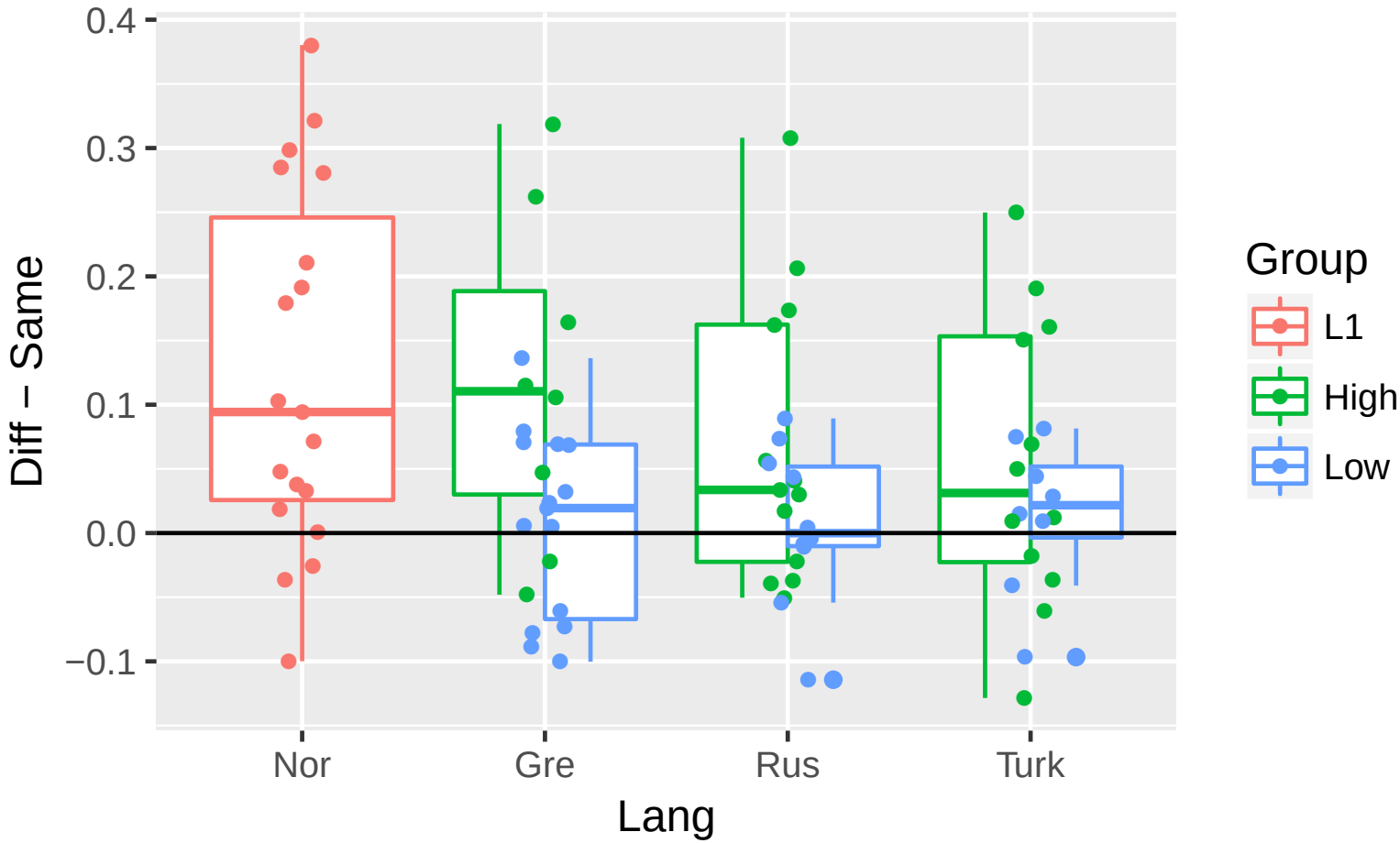
	Early slot (n=10019, groups: Part, 85; Stimulus, 64)			Late slot (n=10379, groups: Part, 85; Stimulus, 64)		
	ß	*Std.err*	*p*	ß	*Std.err*	*p*
1. Intercept (Nor, Same)	-1.193	0.193	5.2e-10 ***	-0.779	0.208	0.0001 ***
2. CondDiff	0.451	0.131	0.0006 ***	0.703	0.148	2.1e-06 ***
3. GroupGreAdv	0.001	0.283	0.9979	-0.038	0.374	0.919
4. GroupGreHigh-inter.	0.686	0.283	0.0153 *	0.203	0.307	0.508
5. GroupRusAdv	0.937	0.293	0.0014 **	1.092	0.318	0.0005 ***
6. GroupRus High-inter.	0.901	0.318	0.0046 **	0.657	0.345	0.056 .
7. GroupTurkAdv	0.394	0.302	0.1911	0.178	0.326	0.583
8. GroupTurkHigh-inter.	1.110	0.341	0.0011 **	0.834	0.370	0.024 *
9. CondDiff: GroupGreAdv	0.037	0.225	0.87	-0.044	0.259	0.864
10. CondDiff: GroupGre High-inter.	-0.50	0.182	0.0059 **	-0.703	0.213	0.0009 ***
11. CondDiff: GroupRusAdv	-0.06	0.186	0.7457	-0.382	0.221	0.084 .
12. CondDiff: GroupRusHigh-inter.	-0.488	0.20	0.0149 *	-0.666	0.237	0.005 **
13. CondDiff: GroupTurkAdv	-0.429	0.193	0.0263 *	-0.386	0.225	0.086 .
14. CondDiff: GroupTurkHigh-inter.	-0.414	0.212	0.0512 .	-0.627	0.252	0.013 *

Effect of Cond by group

Group	Early RoI, looks to target		Late RoI, looks to target	
	Diff-Same	*p*	Diff-Same	*p*
Nor control	0.32 - 0.23	<0.01	0.48 - 0.31	<0.001
GroupGreAdv	0.33 - 0.23	<0.01	0.46 - 0.31	<0.01
GroupGreHigh-inter.	0.36 - 0.37	0.72	0.36 - 0.36	0.993
GroupRusAdv	0.53 - 0.44	0.05	0.65 - 0.58	0.056
GroupRusHigh-inter.	0.42 - 043	0.81	0.48 - 0.47	0.84
GroupTurkAdv	0.31 - 0.31	0.88	0.43 - 0.35	0.067
GroupTurkHigh-inter.	0.49 - 0.48	0.83	0.53 - 0.51	0.711

- But this still doesn't show that e.g. GreHigh behaves significantly different from GreLow or TurkHigh.
- Why can't we just infer group differences from presence/absence of effect within the individual groups?

Variation between participants



Difference between Experimental (diff) and Control (same) conidtion, Proportion of fixations to target in time span from Early to Late Roi.

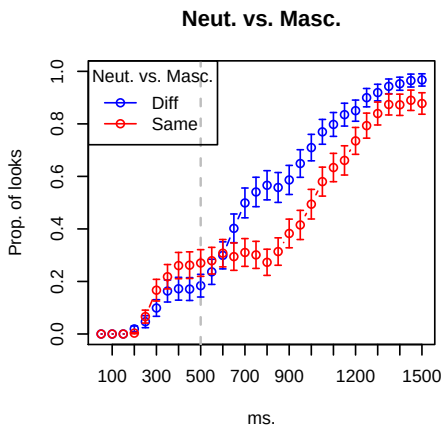
Summary/take home message/More interactions

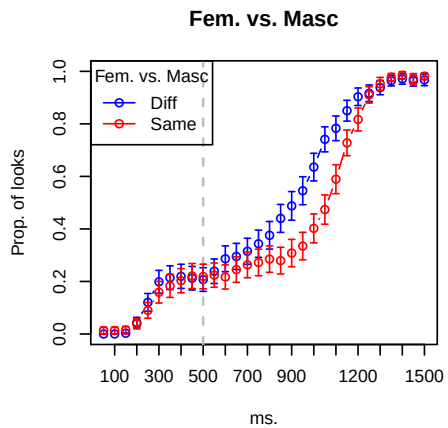
- Simple research questions quickly translate into quite complex paradigms, which requires a large number of participants and items.
- In many "simple" research questions, you are actually looking for a three-way interaction (which is hard to find, ever)
- Example 1: Is there an effect of gender congruency, and if it interacts with proficiency: Cond1 (same/diff) *Cond2 (cong/incong) * Group
- Is there an effect of gender (neut vs. masc) in addition to cond?

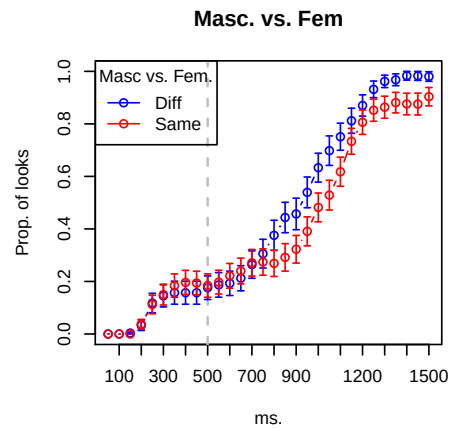
Other dependent measures

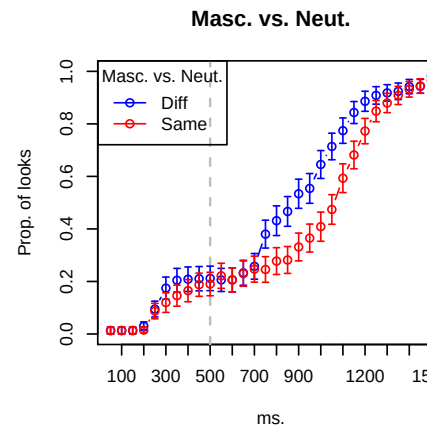
- Average first fixation at target (See Hopp and Lemmerth 2018, Lundquist et al. 2016 etc.): This is an OK measure if you have four or more images on the screen. With two images, early looks to target are not necessarily triggered by sound stimulus.
- Change point estimates, see Dussias et al. 2013.
- Comparing total number or total length of fixations over a longer time span: this may be necessary if timing for whatever reason differ across conditions or items, see Lundquist and Vangsnes 2018.

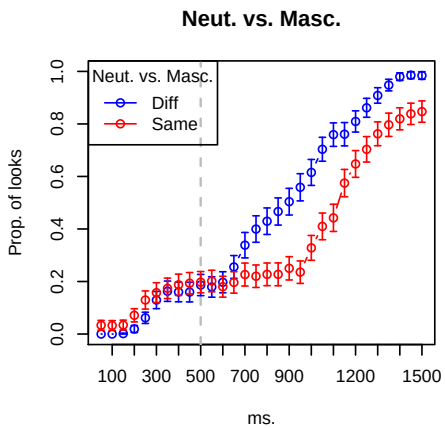
Lundquist and Vangsnes 2018

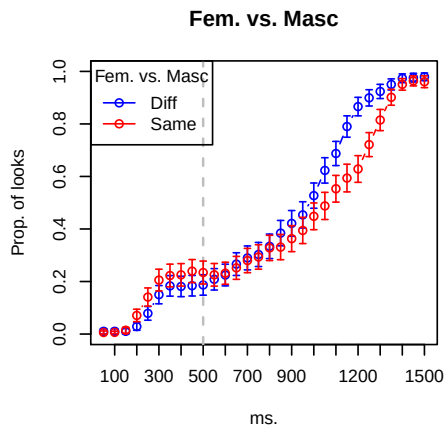


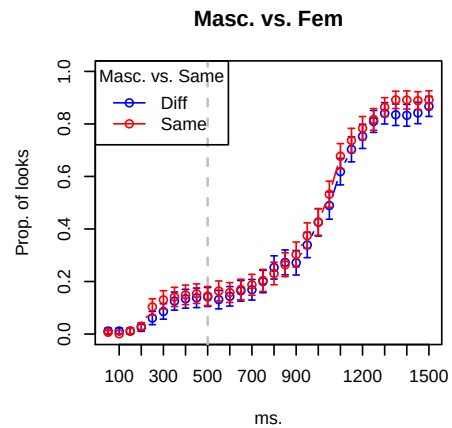


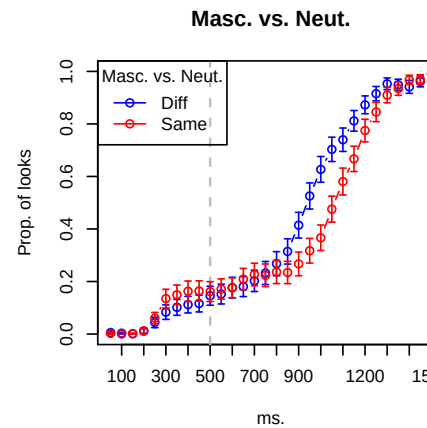




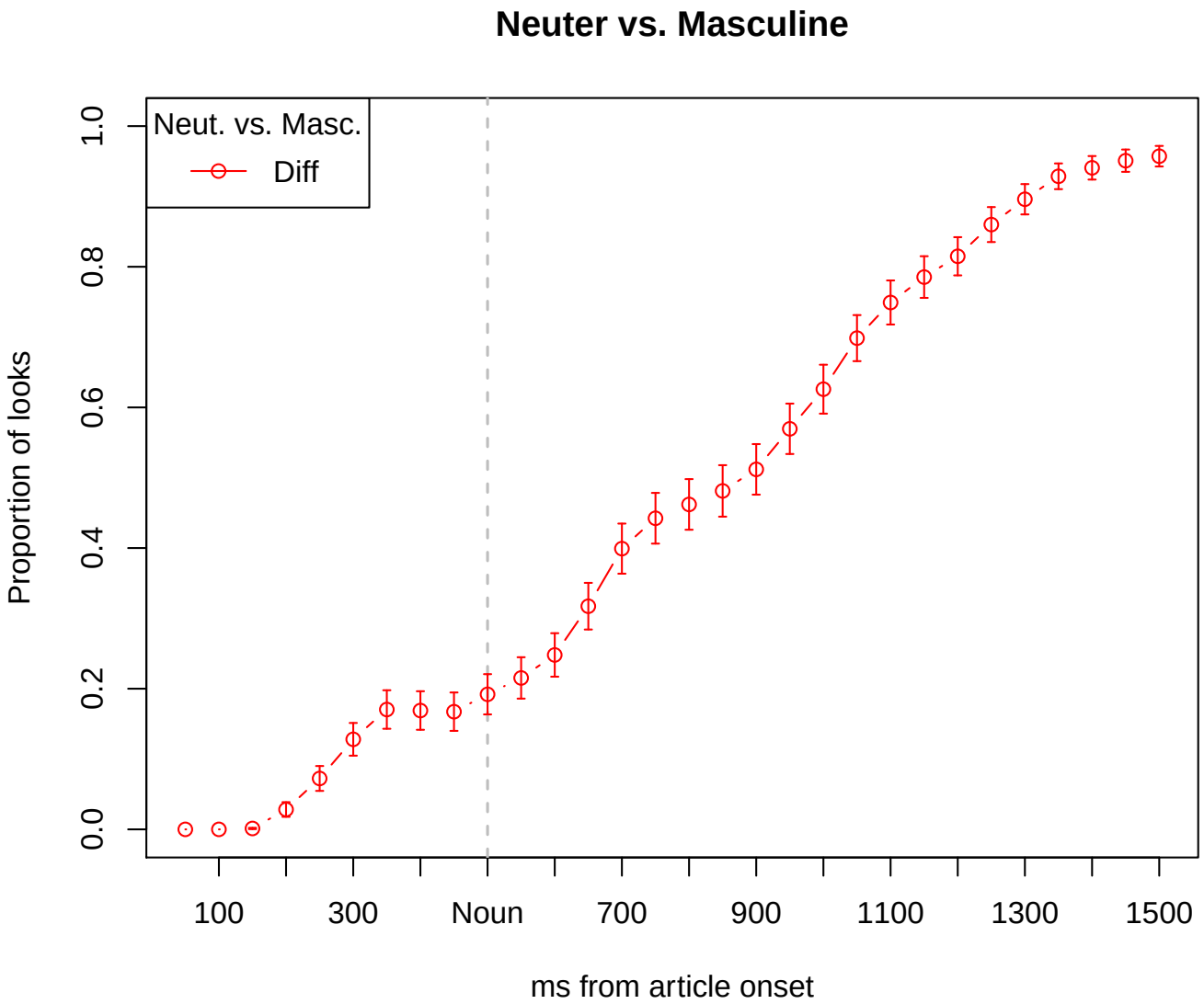




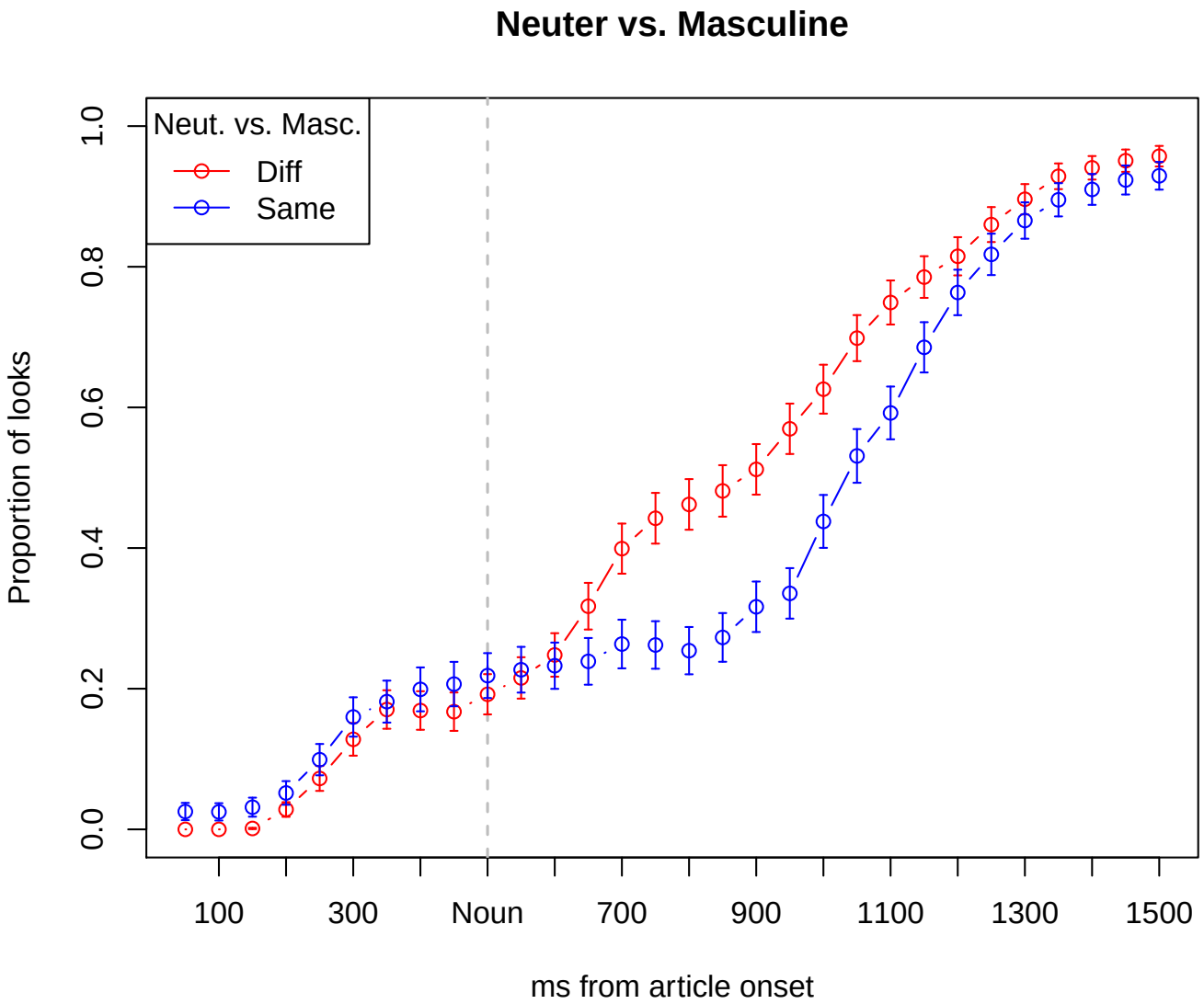




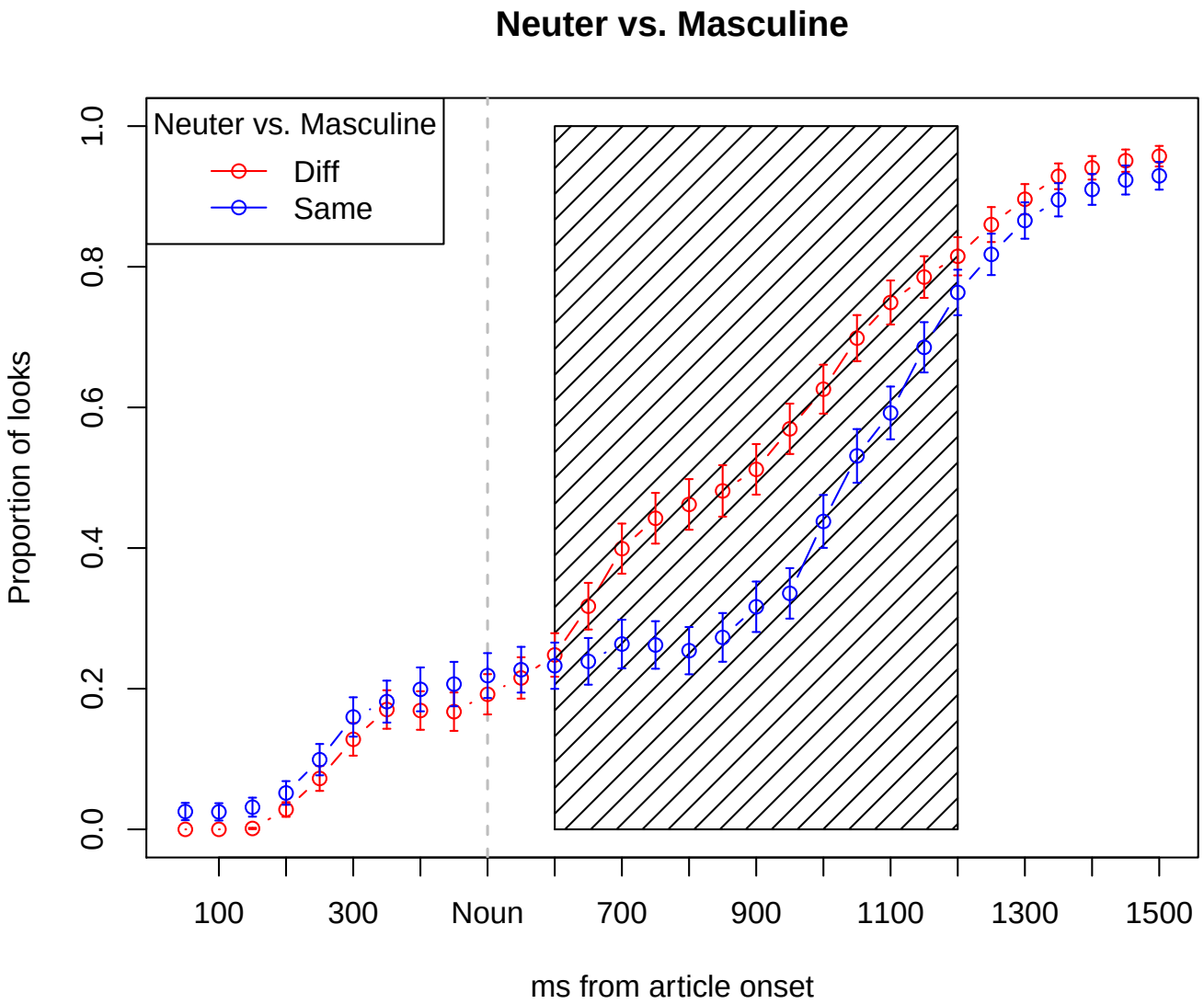
Visualising and analysing the results



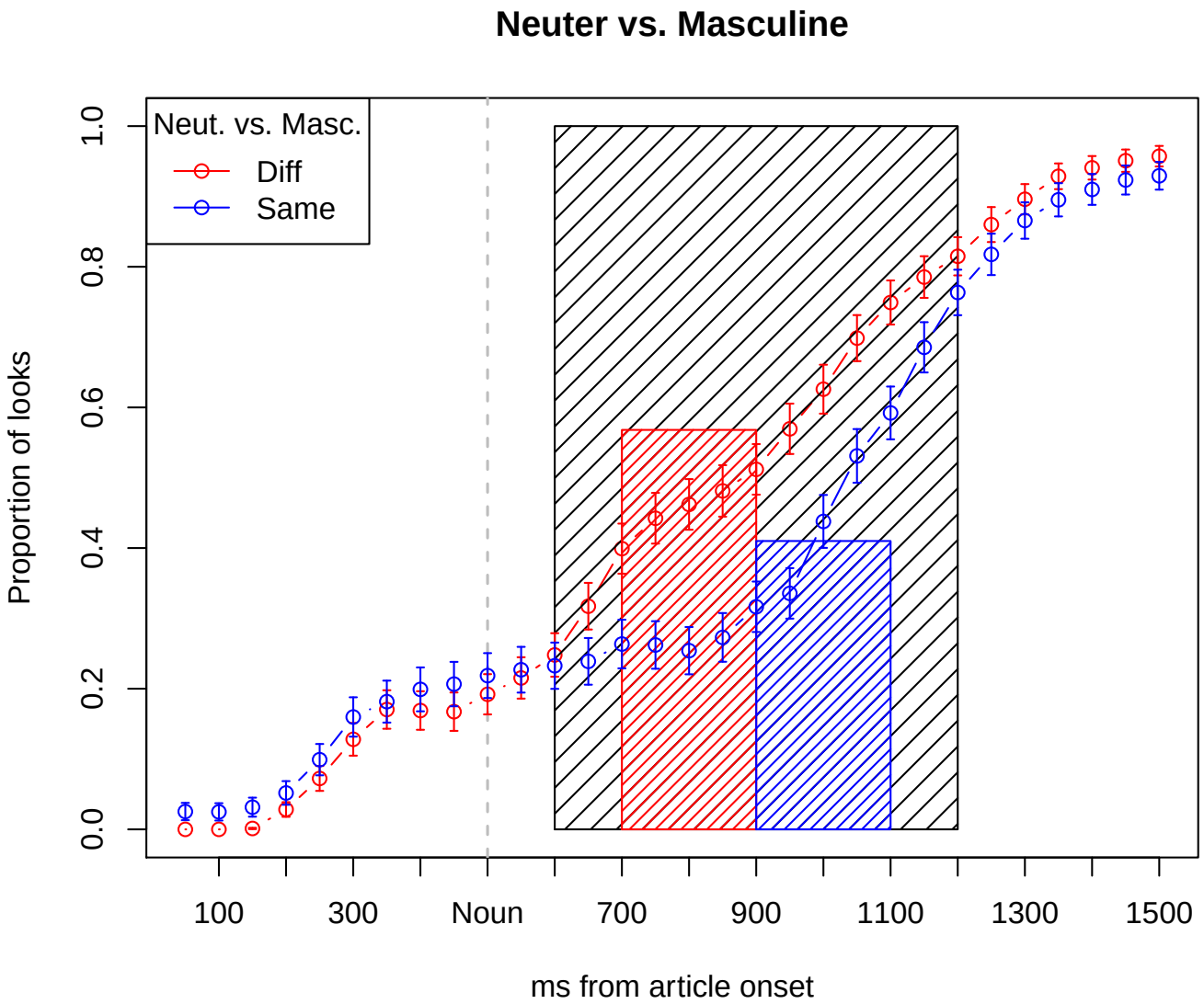
Visualising and analysing the results



Visualising and analysing the results



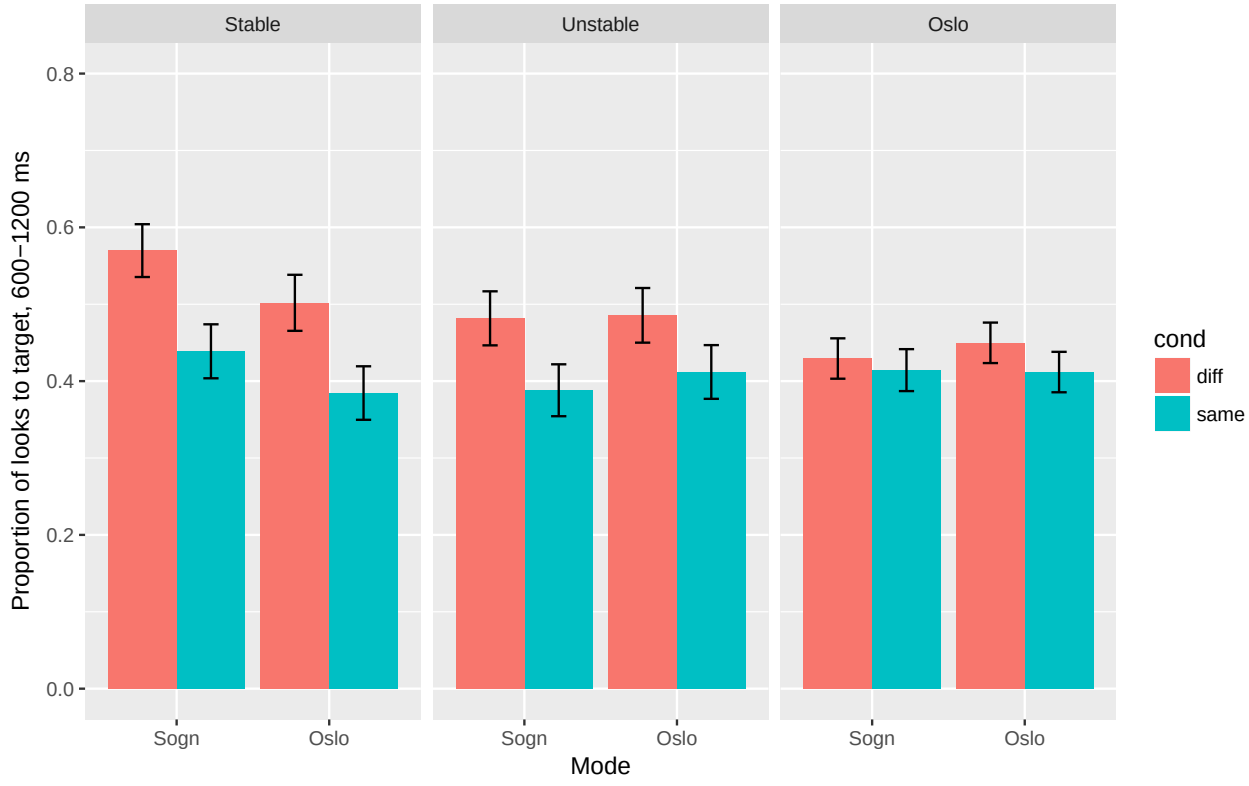
Visualising and analysing the results



	Diff. condition		Same condition	
	Targ.	Dist.	Targ.	Dist.
NeuterMasc	et hus *a house* (N)	en bil *a car* (M)	et hus *a house* (N)	et tåg *a train* (N)
MascNeuter	en bil *a car* (M)	et tåg *a train* (N)	en bil *a bil* (M)	en sykkel *a bike* (M)
FemMasc	ei bok *a book* (F)	en vase *a vase* (F)	ei bok *a book* (F)	ei flaske *a bottle* (M)
MascFem	en trompet *a trumpet* (M)	ei bok *a book* (F)	en trompet *a trumpet* (M)	en vase *vase* (M)

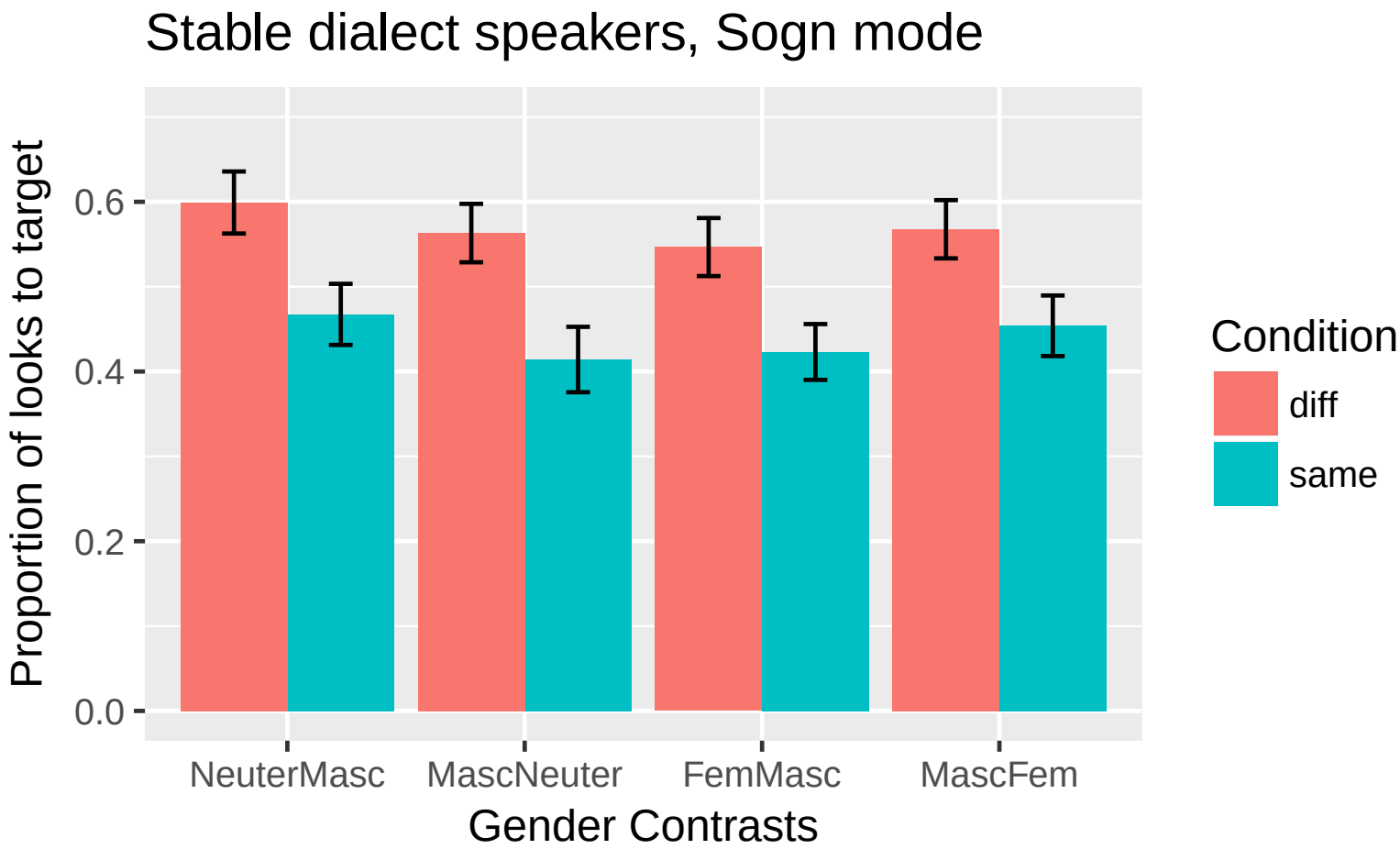
lmer(Prop. of looks to target, 600-1200ms (Dep. variable)
~Diff/Same condition * Gender Contrast (Predictors)
+ (1+Cond|Participant) + (1+Cond|Item) (Random eff.)

Figure: Effects of Condition (Different/test vs. Same/control) and Mode for the three groups. Error bars indicate 95% confidence intervals.



Main effect of Cond ($\chi^2 = 22.708$, $df = 1$, $p < 0.001$), a Group x Cond interaction ($\chi^2 = 23.157$, $df = 4$, $p < 0.001$) and a three-way Cond x Group x Mode interaction ($\chi^2 = 20.361$, $df = 6$, $p < 0.01$).

Stable Sogn: Main effect of Mode ($p < 0.001$)

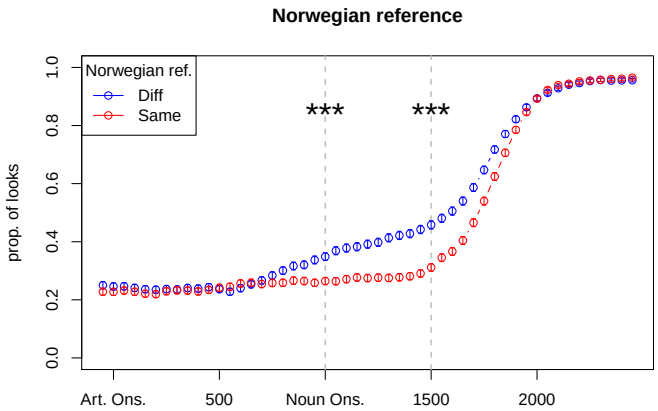


Main effect of same/diff Condition ($p < 0.001$).

Excluding trials based on production, or not?

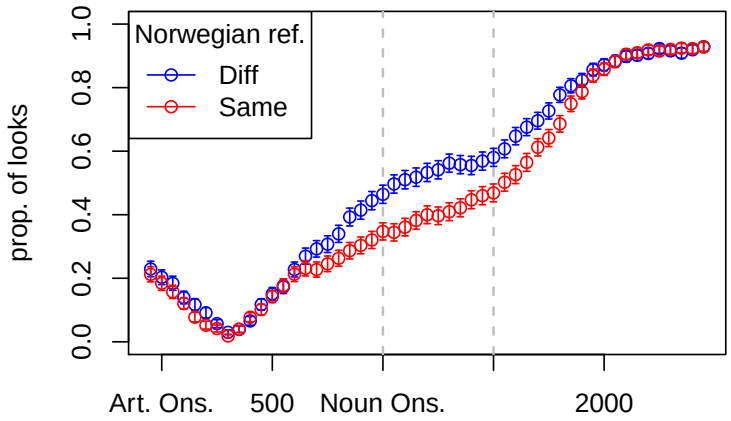
- Hopp and Lemmerth 2018: Every eye tracking trial is preceded by a naming task where the participant names each item with a gender marked article.
- Trials where the crucial items are not correctly named are excluded from the analysis.
- This will be very difficult to implement in a web-based setting: collecting spoken responses is yet not fully trivial, asking participants for written responses might make the experiments very long.
- Other studies will include all trials, and not necessarily check the production. Pros
  1. No assumption about a production/comprehension symmetry.
  2. Often, the same stimulus appears several times throughout the experiment, which means that you might acquire some items over the experiments.
  3. Makes the stats easier. (equally many trials of the different conditions)

Exclude trials based on early fixations

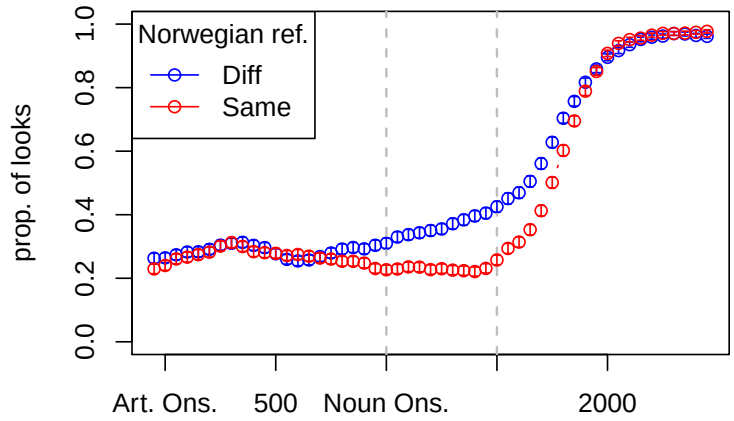


- Looks to target at Article onset is clearly not triggered by the article. Should we exclude them from the analysis?
- If your dependent measure is first fixation, yes, but otherwise no.

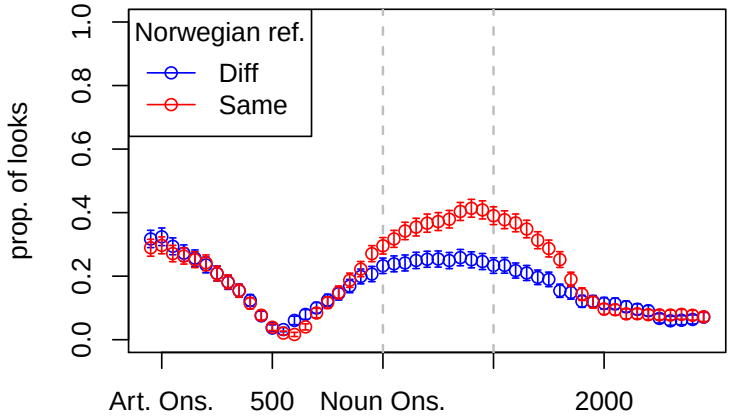
Early look distractor



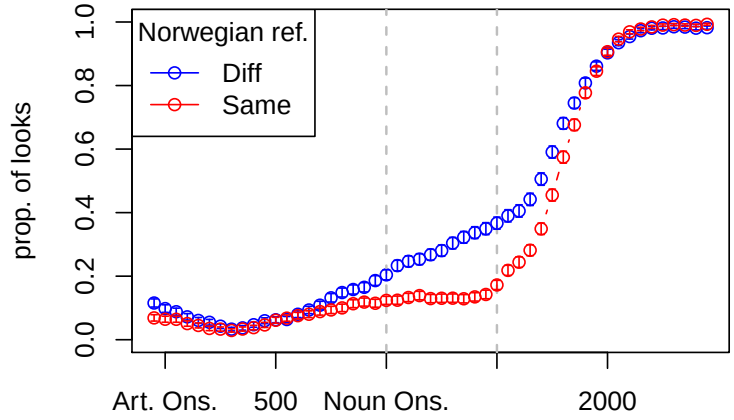
Early looks fill



Looks at dist, when early fix at target



Late saccade from fixation



Even in trials where the target is fixated prior to onset of disambiguating cue (here, gender-marked article), we find an effect of the same-different manipulation! Do not throw away data based on early fixations at the target - you never know what triggers a saccade to an image!

summary

- Lot's of things influence your fixation pattern: grammatical markers is only one of them. Lexical markers more reliable.
- There is plenty of variation between participants – many participants do not show an effect of the manipulation, even in your control group: Make sure to have many participants!
- If you make comparisons between two groups (which you usually will do), make sure that your statistical model actual includes both group.