IN1140

Eksamensverksted

semantikk syntaktiske trær * morfemer avledning syntaks
fonetikk flertydighet regex VP innholdsord AdvP funksjonsord
lemmatisering stemming korpus naive bayes AdjP betinget
sannsynlighet ? Isolerende språk typologi suffiks lingvistikk
statistisk synonymi språkmodell affiks sentimentanalyse
produktesetningen data sparseness n-gram regelbasert tagger
parsing form features fe Språkteknologi morfologi
fonologi bigram frekvenser maximum likelihood estimation
chunking konteksfri features grammatikk agent NP
dependensgrammatikkkrsiv maskinlæring informasjonsekstraksjon
pragmatikk antonymi setningssemantikk leksikal semantikk
priorsannsynligheten PP oversettelse IDF-vekting homonymi

EKSAMENDAGSPLAN

- Sov godt og lenge
- 14:30 – Dørene i Silurveien 2 åpnes
  - Finn riktig sal (i StudentWeb)
  - Huskeliste: Legitimasjon, skrivesaker, legge telefon og klokke i sekken og UiO-passordet deres
- 15:00 – Eksamenstart
- 19:00 – Eksamenslutt



Informasjon

Eksamen i IN1140

Introduksjon til språkteknologi

25 november kl. 09:00 - 13:30 (4,5 timer)

Eksamenssettet inneholder 12 oppgaver.
Alle hjelpemidler er tillatt (lærebok, nettressurser, notater osv.) Det er ikke tillatt å samarbeide eller kommunisere med andre under eksamen om oppgavene.
Oppgaven har én oppgave som krever filopplasting.
httlps://www.uio.no/studier/eksamen/innlevering/alternativer-for-handtegninger.htm

Man kan trekkes ut til samtale for å kontrollere eierskap til sin besvarelse
https://www.mn.uio.no/om/hms/koronavirus/kontrollsamtale/
Samtalen har ikke innvirkning på sensuren/karakteren, men kan lede til at instituttet oppretter fuskesak. Les mer om hva som regnes som fusk på UiOs nettsider:
https://www.uio.no/om/regelverk/studier/studier-eksamener/fuskesaker/

For øvrig gjelder informasjonen på nettsiden om eksamensavvikling ved MN høsten 2020:
https://www.mn.uio.no/om/hms/koronavirus/eksamen-2020.html

Kontaktinfo:
https://www.mn.uio.no/om/hms/koronavirus/brukerstotte/brukerstotte-eksamen-h20.html

1 Regulære uttrykk (15 poeng)

Du vil kjøpe en togreise fra Oslo S til Bodø. Du er fleksibel når det gjelder dato, så lenge det er før (men ikke inkludert) 01.02.2021. Du vil ha en billett som koster maks 1999kr, og vil reise fra Oslo S tidligst kl. 08:00 og seinest kl. 22:00. Du ønsker også å se prisene for både Standard og Premium billetter.

Skriv et regulært utrykk som fanger resultatene av søket ditt slik at resultatene **alltid** er skrevet i følgende format:

```
ukedag dag måned år\nylinjeprisklasse tid billettype
```

Merk følgende formateringer:

- dag skal være melllom **01** og **31**.
- måned skal være **november**, **desember**, og **januar**.
- år skal være **2020** eller **2021**.
- prisene skal være mellom **1** og **1999** kroner. Merk at prisen kan bli skrevet som **kr** eller **nok**
- klokkeslett skal alltid starte med **kl.** og være mellom **08** og **22**. Minutter skal være mellom **00** og **59**. Klokkeslett skal være skrevet som **kl.** *mellomrom* time:minutter.
- billettype skal være enten **Standard** eller **Premium**.
- det er *mellomrom* mellom hver element, utenom etter *år*, der skal det være *ny linje*.

For eksempel skal følgende strenger gjenkjennes av ditt regulære utrykk:

1. `mandag 01 desember 2020299kr kl. 09:30 Standard`

Prøveeksamen

Instructions

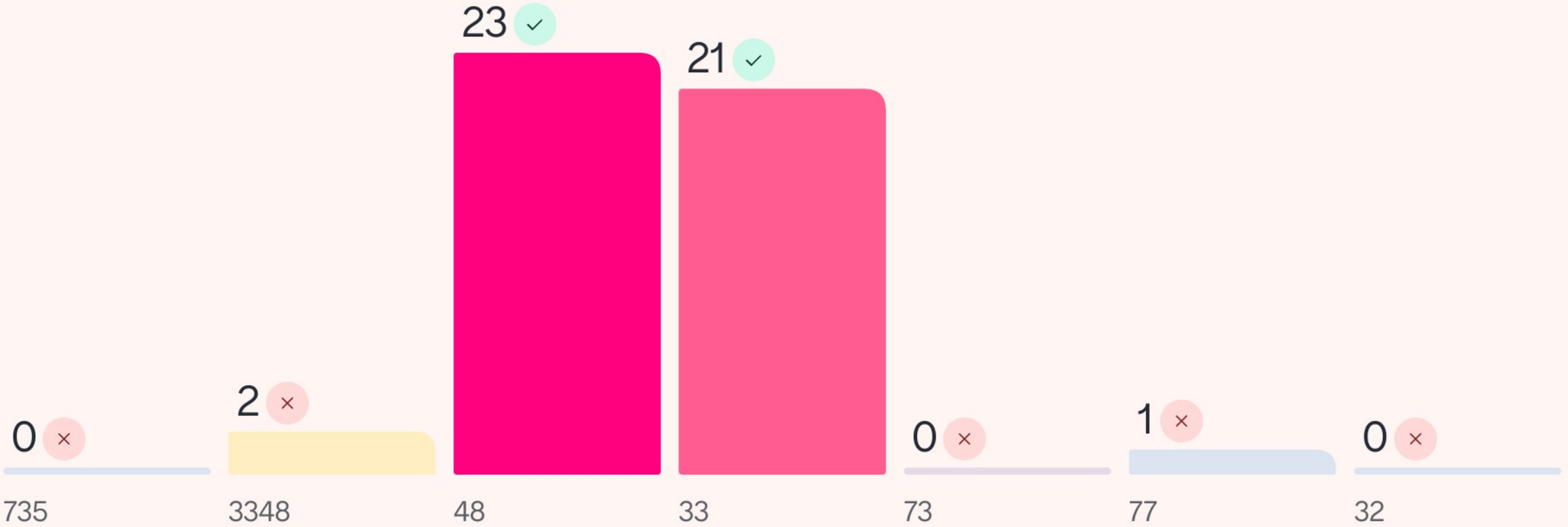


Regulære uttrykk

Vi har følgende regulære uttrykk: /3{2}|48|7{3,5}/

Hvilke to alternativ inneholder tekststrenger der hele strengen vil matches av det uttrykket?
(Alle alternativene er tekststrenger)

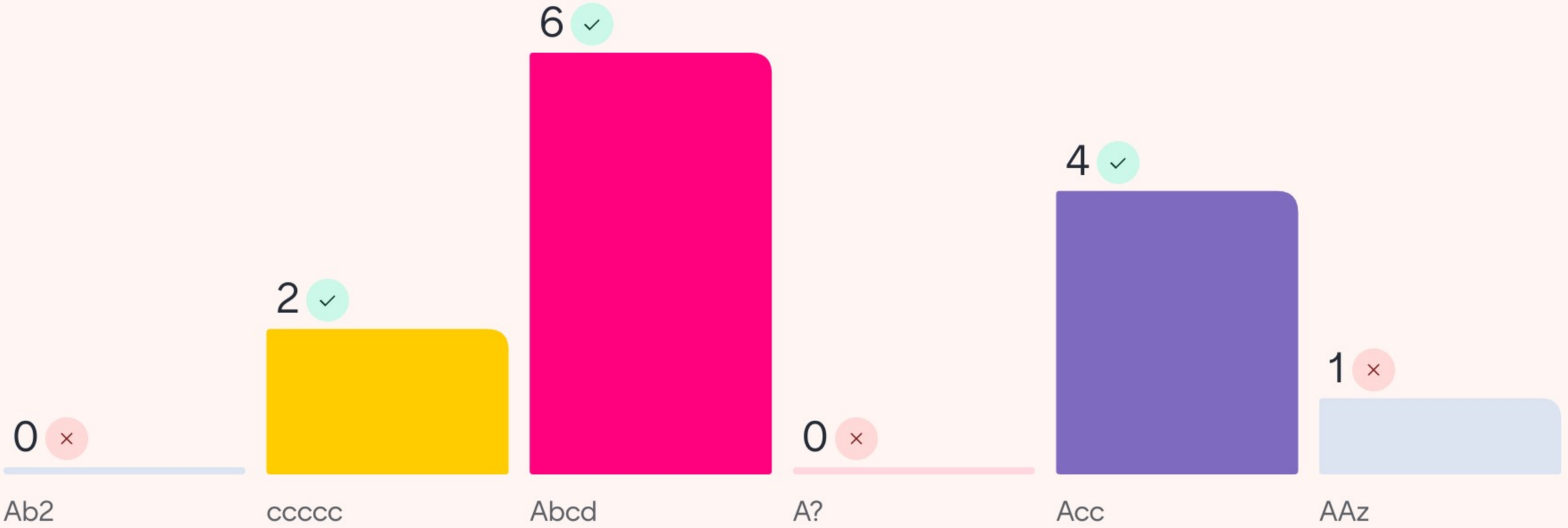
Vårt regulære uttrykk: /3{2}|48|7{3,5}/



Vi har følgende regulære uttrykk: /A?[a-z]+/

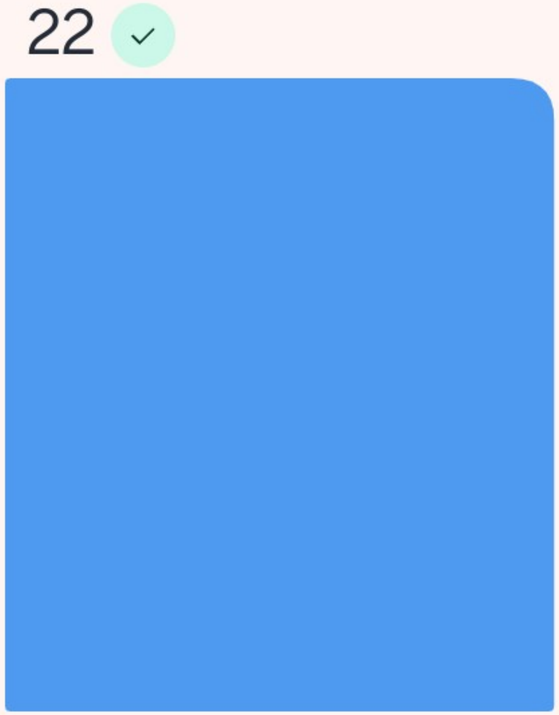
Hvilke tre alternativ under inneholder tekststrenger der hele strengen vil matches i én match av det regulære uttrykket?
(Alle alternativene er tekststrenger)

Vårt regulære uttrykk: /A?[a-z]+/



Representasjon av mellomrom: Hvilket alternativ inneholder et regulært uttrykk som matcher "g x" ?

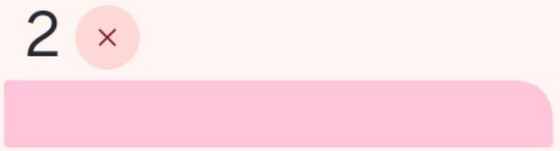
Hvilket alternativ inneholder et regulært uttrykk som matcher "g x" ?



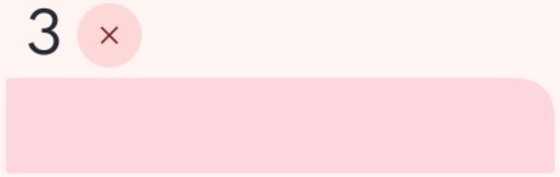
Mellomrom representeres med mellomrom: /g x/



Mellomrom representeres med skråstrek foran underskrekning: /g_x/



Regulære uttrykk kan ikke matche mellomrom



Mellomrom representeres med krølleparentes med et mellomrom inni: /g{ }x/

Datoformater

Korrekte norske datoangivelser som ikke angir måned med bokstaver, er i følge Språkrådet de følgende:

5.6.2014 05.06.2014 5.6.14 05.06.14

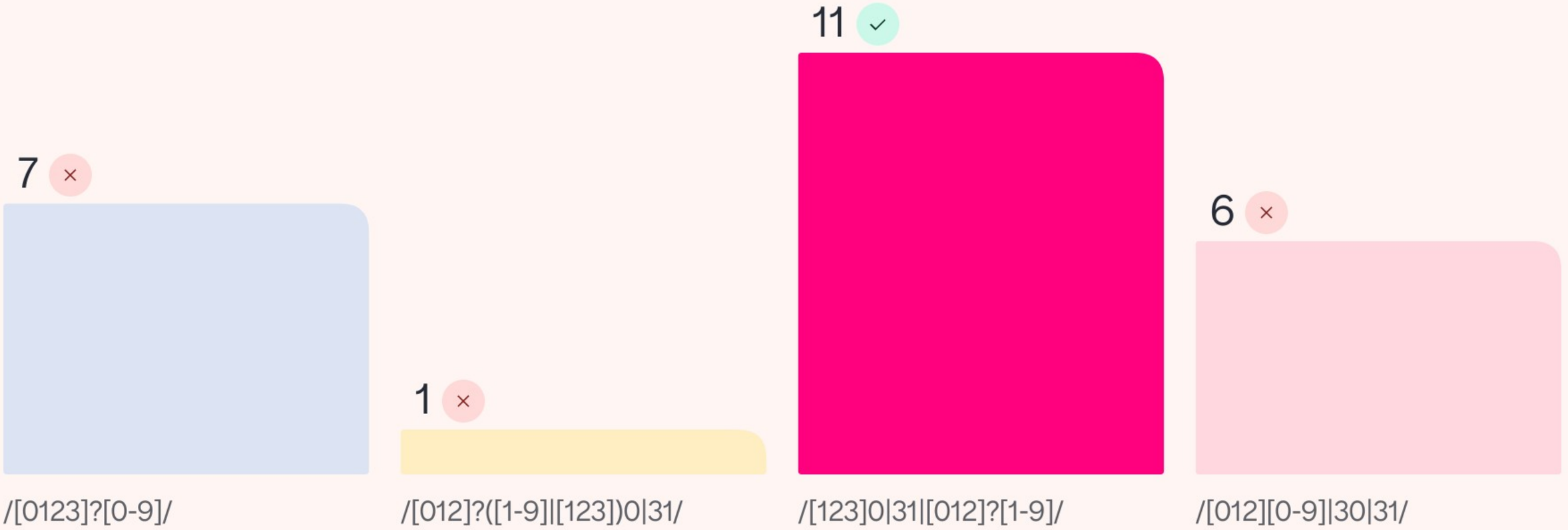
(Alle i rekkefølgen dag-måned-år)

Vi skal finne regulære uttrykk som matcher hver sin del av korrekte datoangivelser. For hver del av tekststrengen vi ser på, skal vi finne regulære uttrykk som kun hvis formatet er korrekt, matcher hele denne delstrengen.

Deloppgave 1: Dag

Dag skal angis med et tall fra og med 1 til og med 31. Ensifrede tall kan ha 0 foran.
Hvilket regulære uttrykk oppfyller de ovenfor nevnte kravene for angivelse av dag i måneden?
Hvis du finner flere korrekte, velger du det korteste.

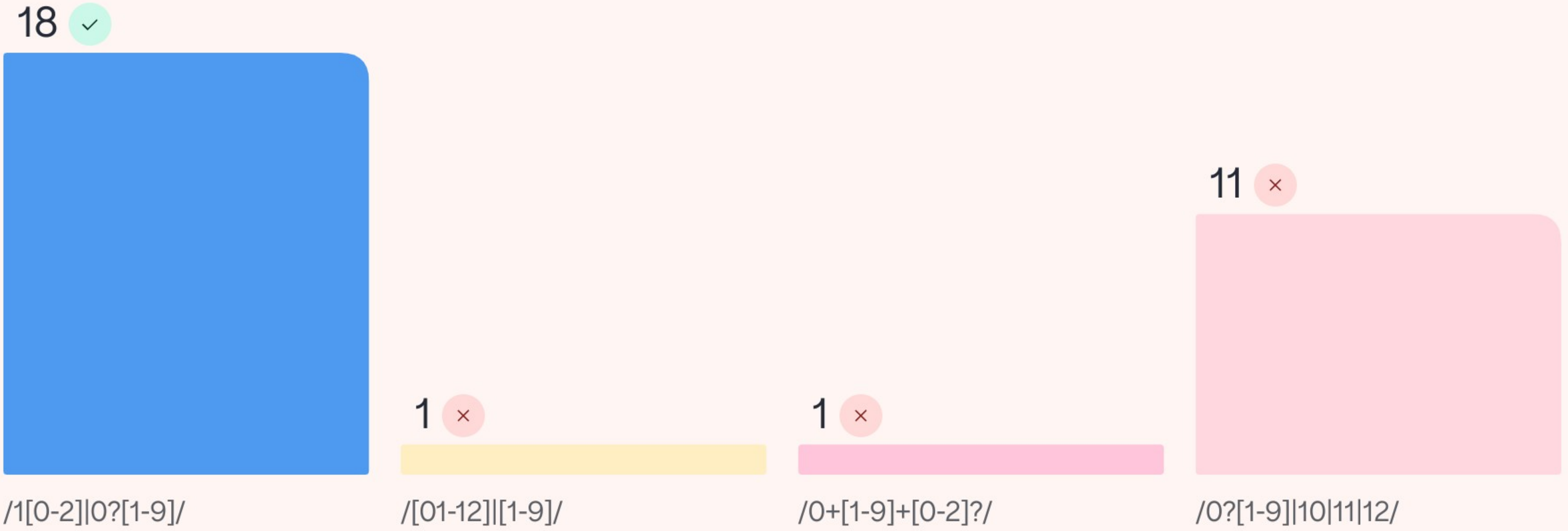
Deloppgave 1: Dag



Deloppgave 2: Måned

Måned skal være et tall fra og med 1 til og med 12. Ensifrede tall kan ha 0 foran.
Hvliket regulære uttrykk oppfyller de ovenfor nevnte kravene for angivelse av måned?
Hvis du finner flere korrekte, velger du det korteste.

Deloppgave 2: Måned



Ordklasser

I IN1140 har vi primært undersøkt tre ulike kriterer for ordklasseinndeling.
Redegjør kort for hvilke kriterier det er snakk om og hvordan de brukes.
Det holder med 2-3 setninger for hvert kriterie, kom gjerne med eksempler.

Diskuter med de ved siden av dere!
Dere har 4 minutter!

Ordklasser

Løsningsforslag:

1. Formelle eller morfologiske kriterier: Hvilke bøyningsformer har ordet?

F.eks: hare - haren og redd - reddere

*harere og *redden

2. Funksjonelle eller syntaktiske kriterier: Hvordan kan ordet kombineres med andre ord?

en hare, en redd hare og redd for ilden,

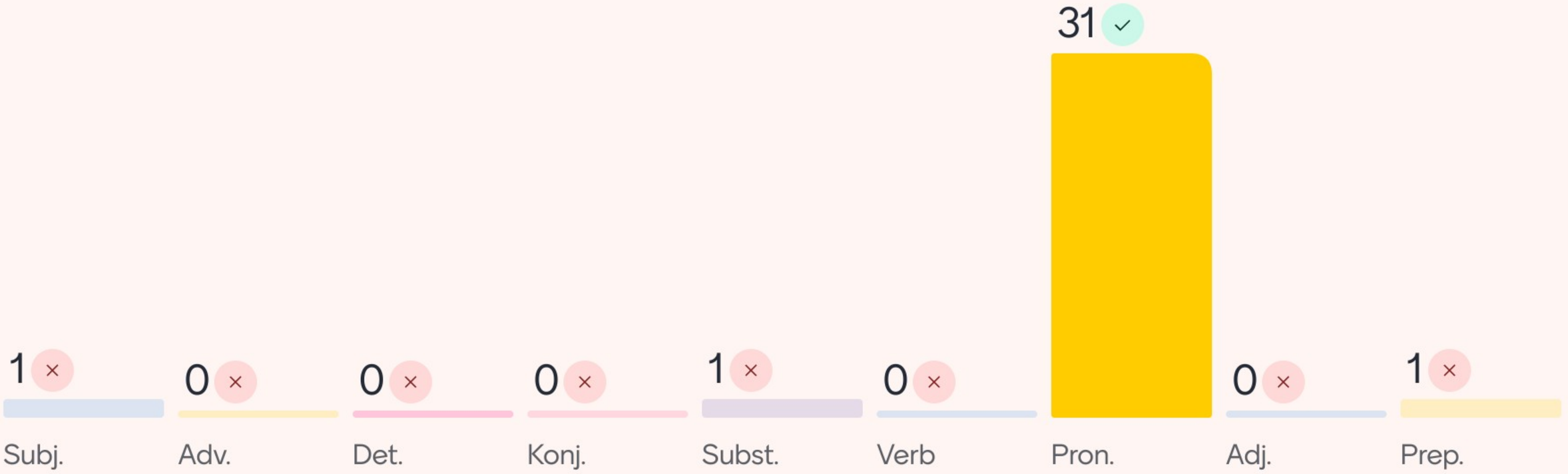
*en redd og *hare for ilden

3. Betydningsmessige eller semantiske kriterier: Hva er typiske betydninger hos ord i ordklassen?

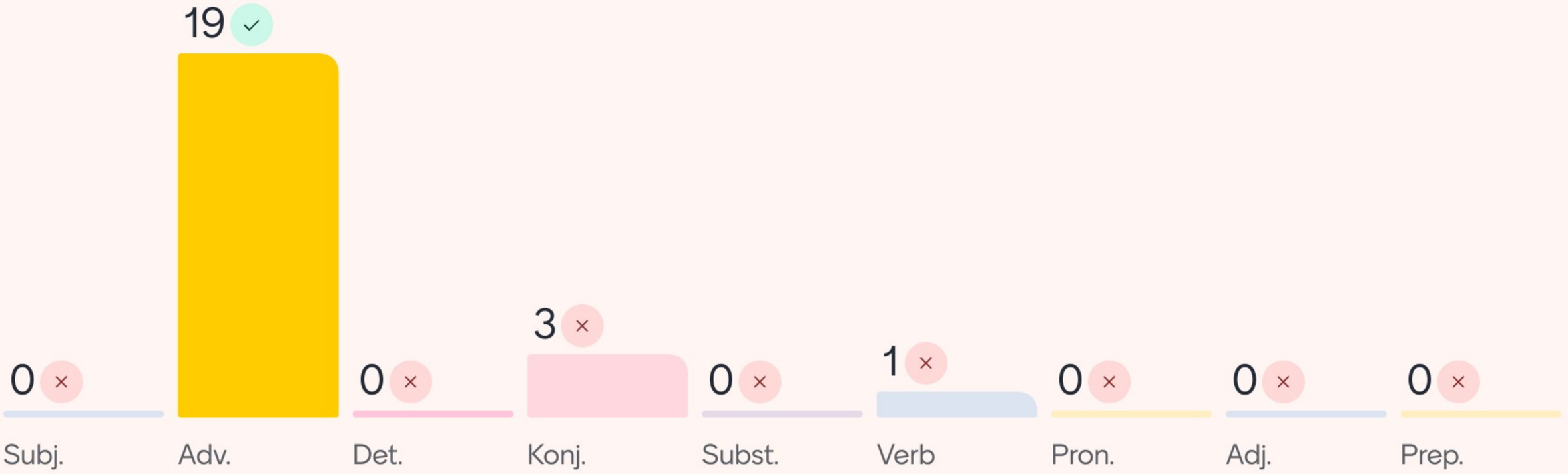
hare - dyr, levende vesen, redd - egenskap

4

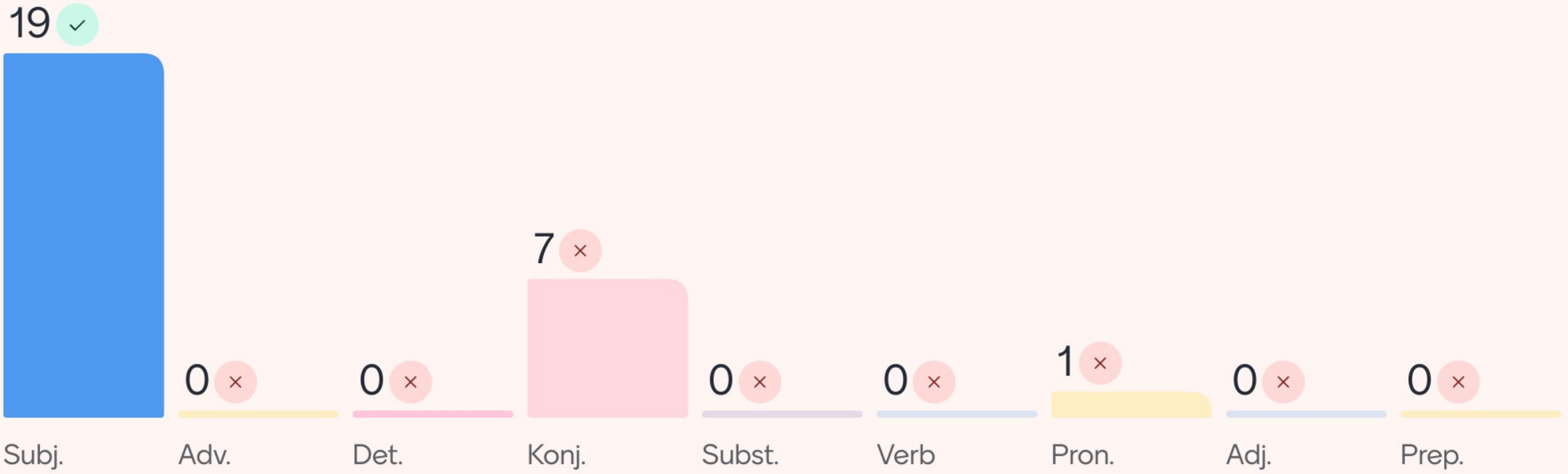
"**Jeg** gjør det kanskje hvis du ber meg om det, men det store problemet er om du stoler på meg eller ikke"



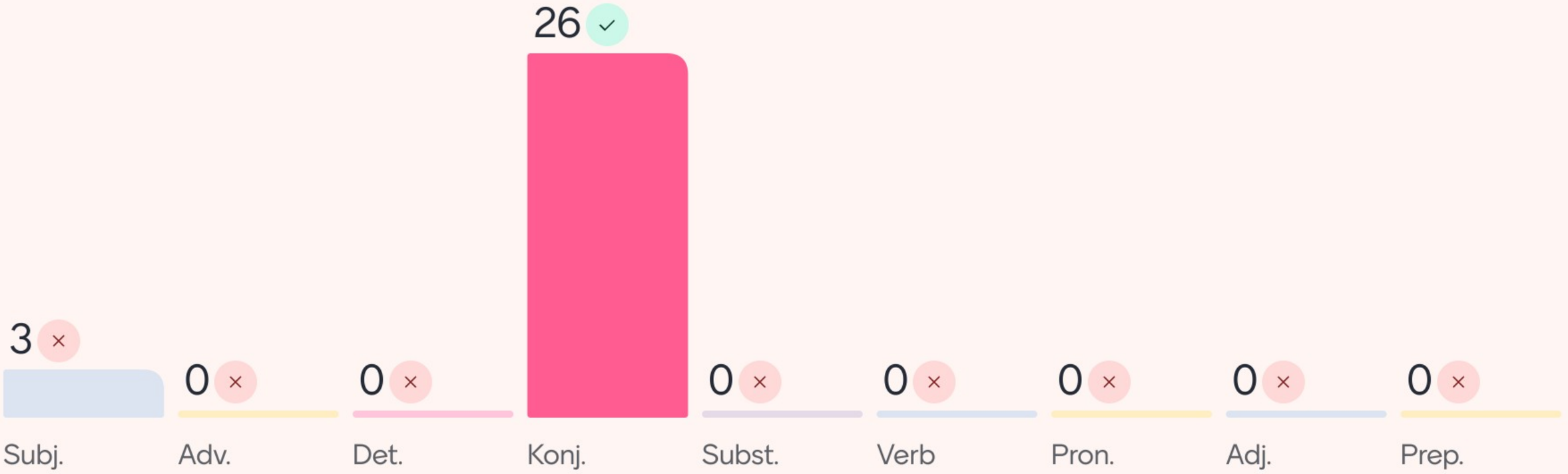
"Jeg gjør det **kanskje** hvis du ber meg om det, men det store problemet er om du stoler på meg eller ikke"



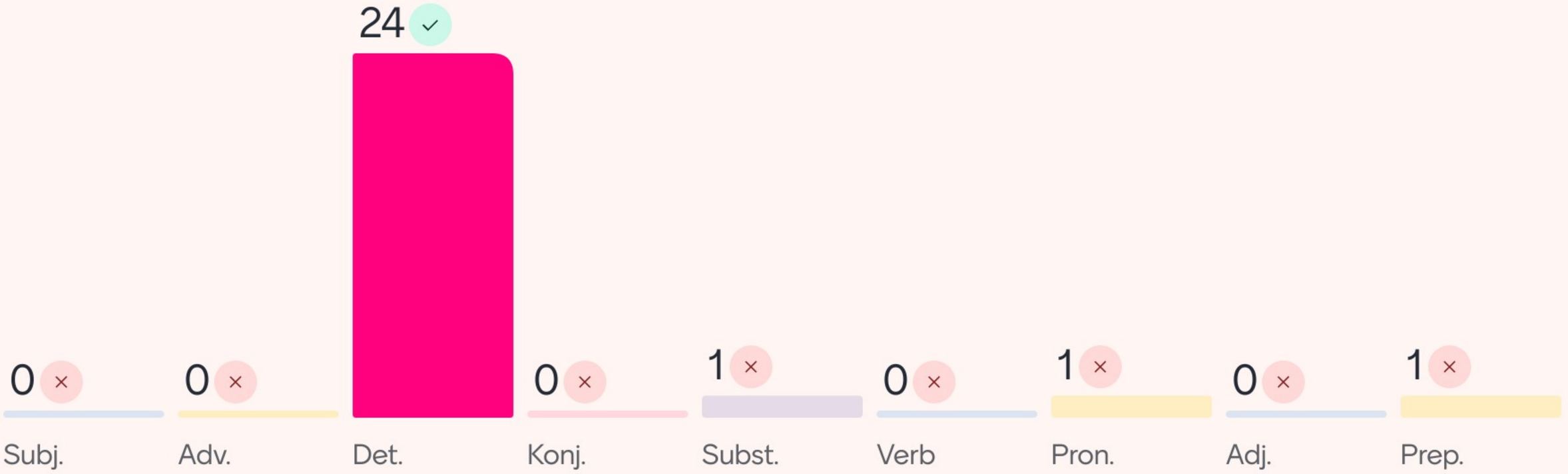
"Jeg gjør det kanskje **hvis** du ber meg om det, men det store problemet er om du stoler på meg eller ikke"



"Jeg gjør det kanskje hvis du ber meg om det, **men** det store problemet er om du stoler på meg eller ikke"

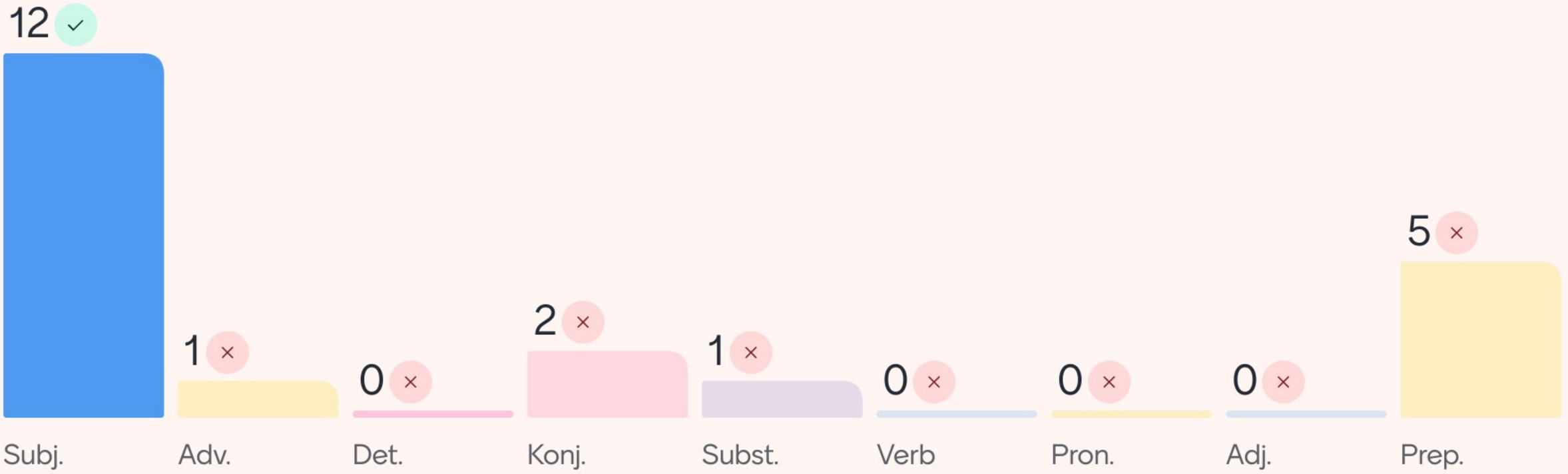


"Jeg gjør det kanskje hvis du ber meg om det, men **det** store problemet er om du stoler på meg eller ikke"

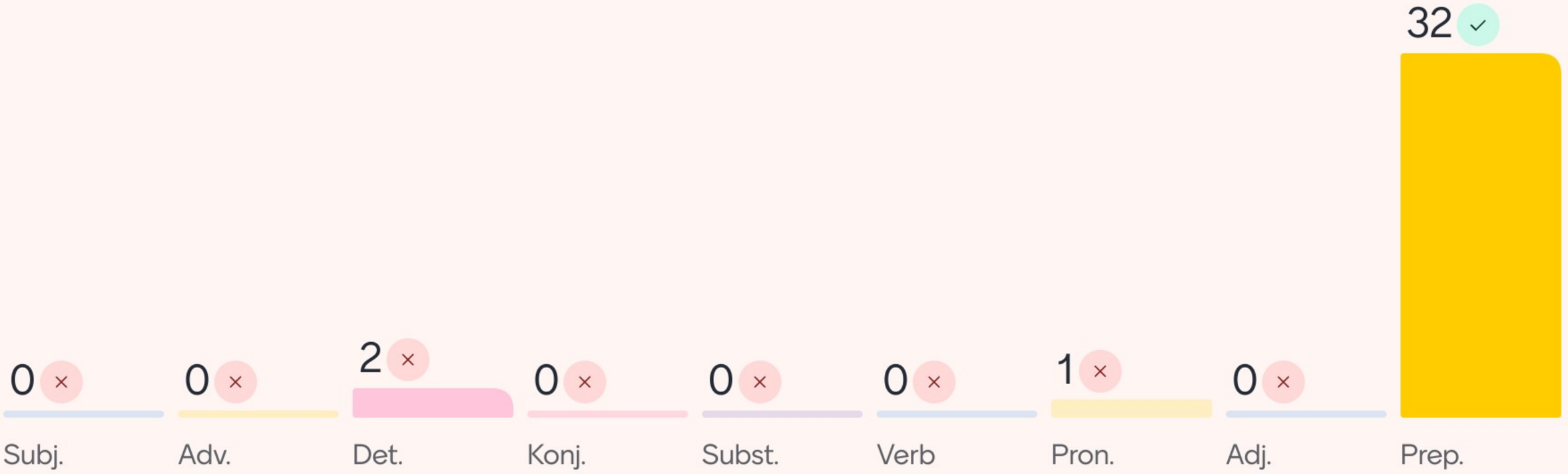


3 27

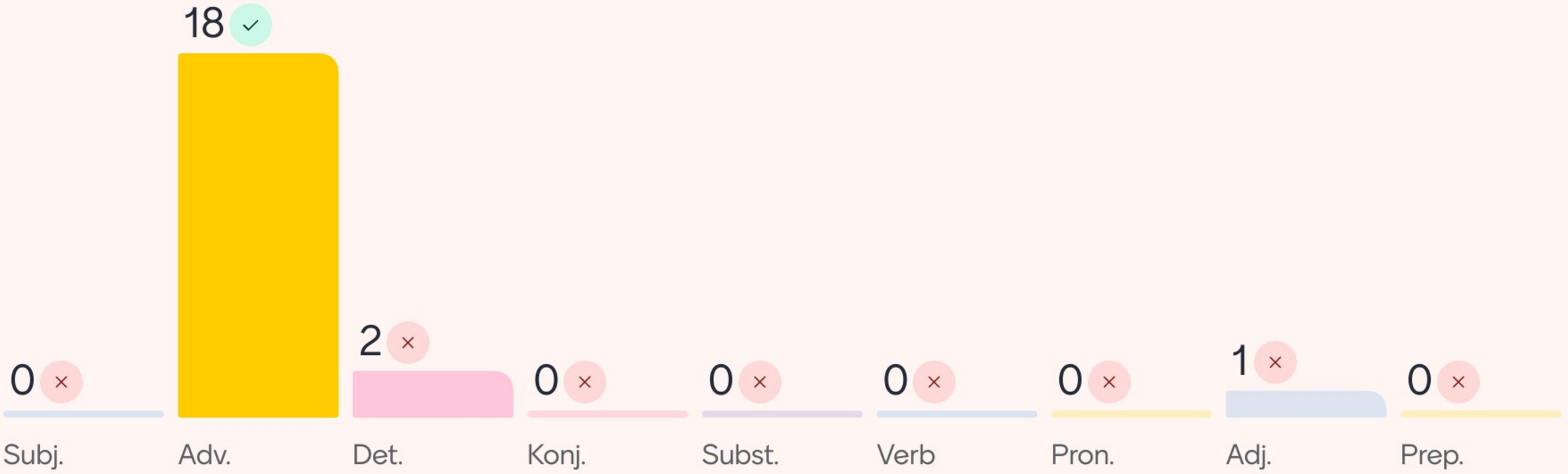
"Jeg gjør det kanskje hvis du ber meg om det, men det store problemet er **om** du stoler på meg eller ikke"



"Jeg gjør det kanskje hvis du ber meg om det, men det store problemet er om du stoler **på** meg eller ikke"



"Jeg gjør det kanskje hvis du ber meg om det, men det store problemet er om du stoler på meg eller **ikke**"

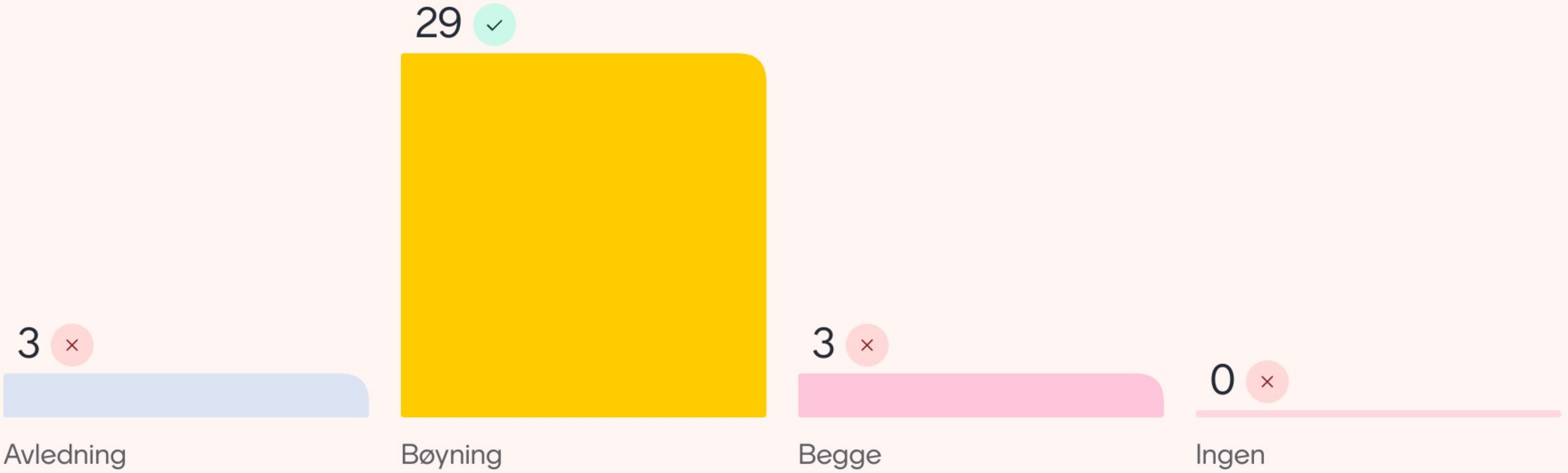


Orddannelse

For hvert av ordene skal du avgjøre om de er dannet ved bøyning, avledning, begge deler eller ingen av delene.

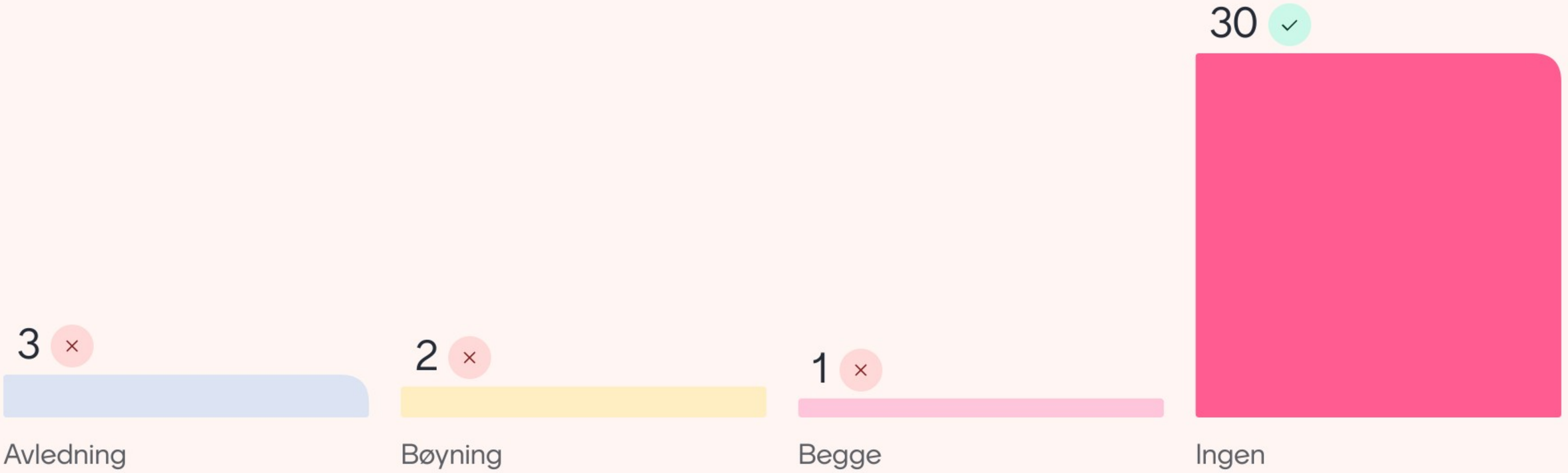
5

kuler



2
35

strømpe



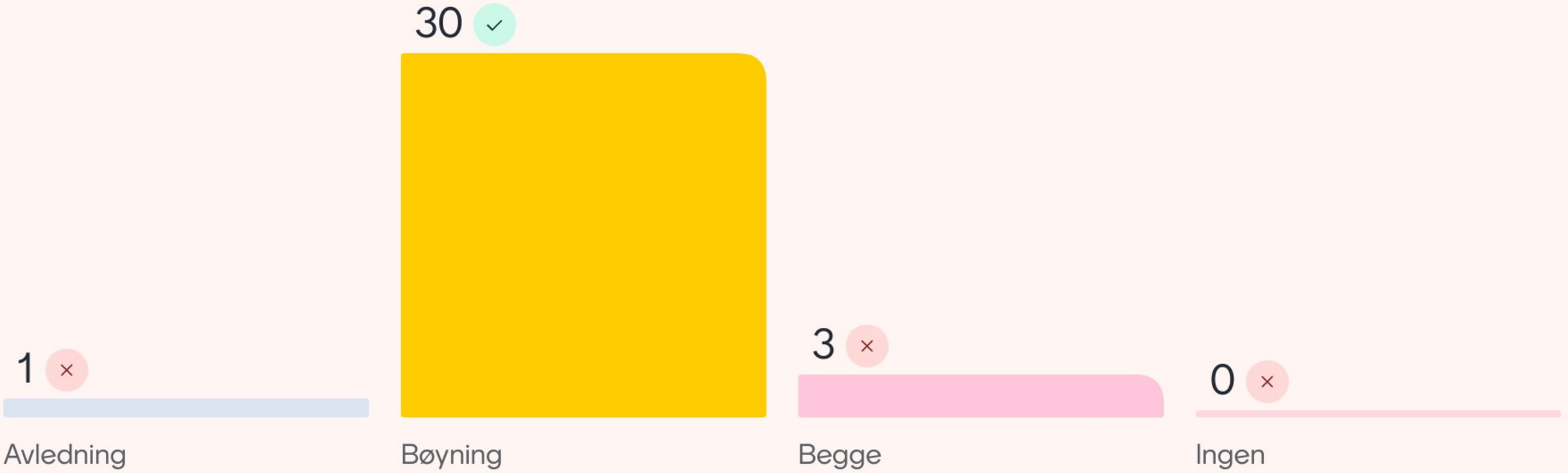
3 36

fest



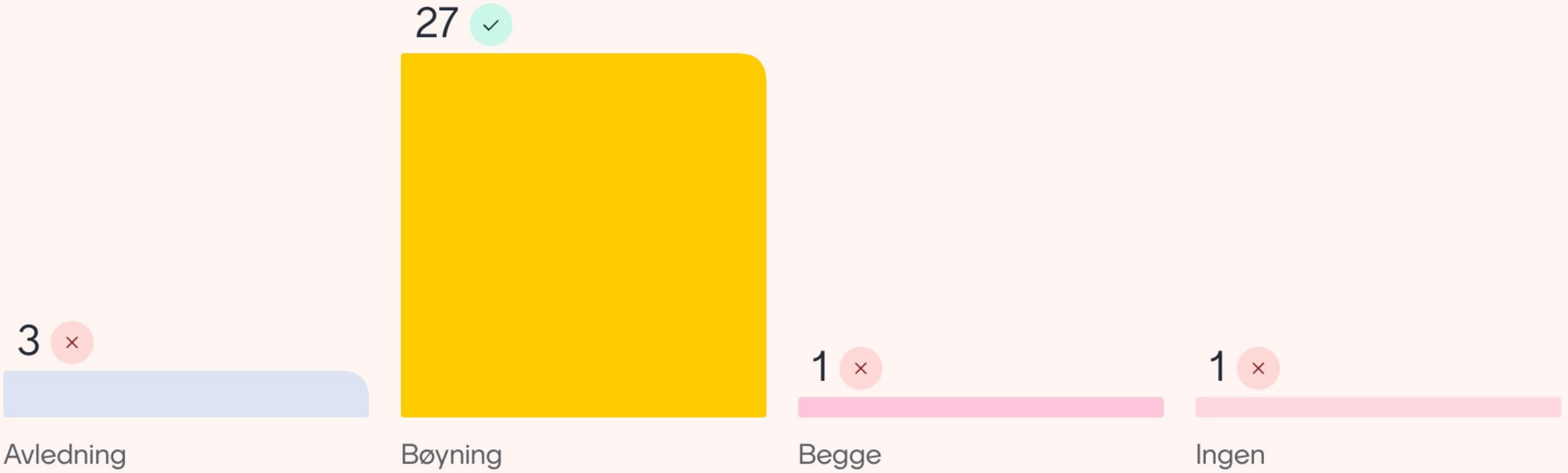
2 31

danset



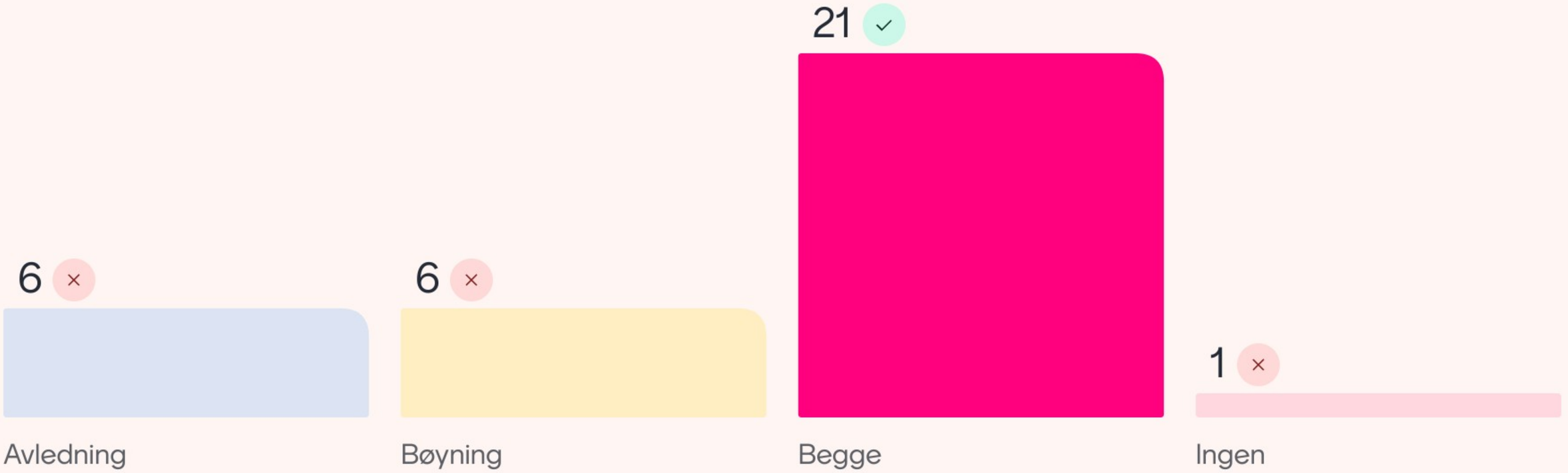
2
34

julen



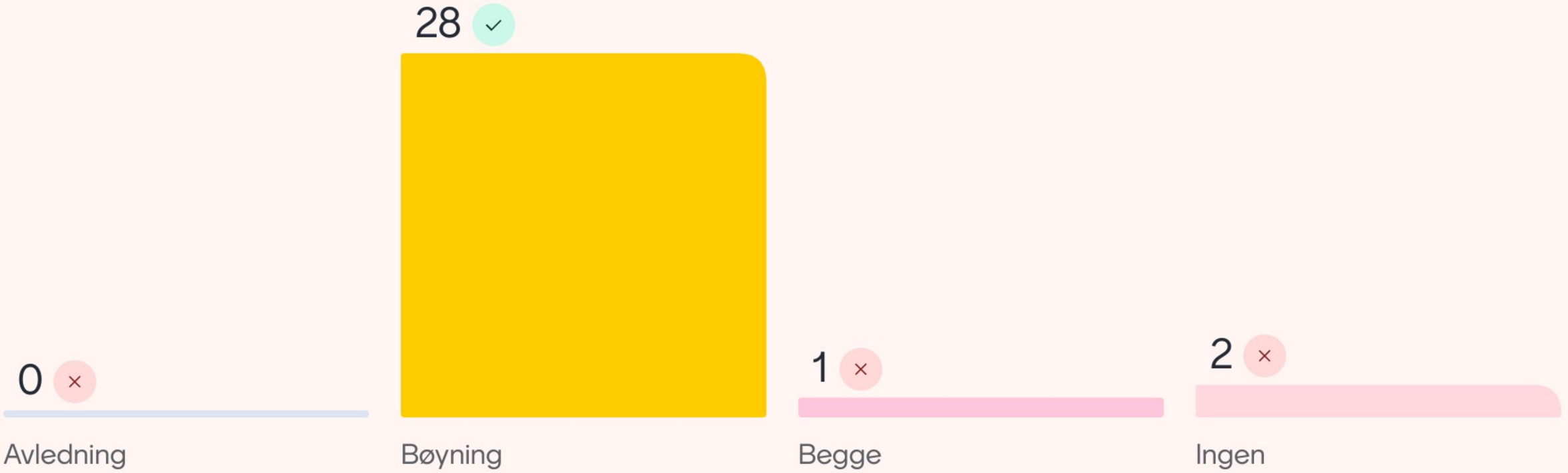
1
32

hjertelige



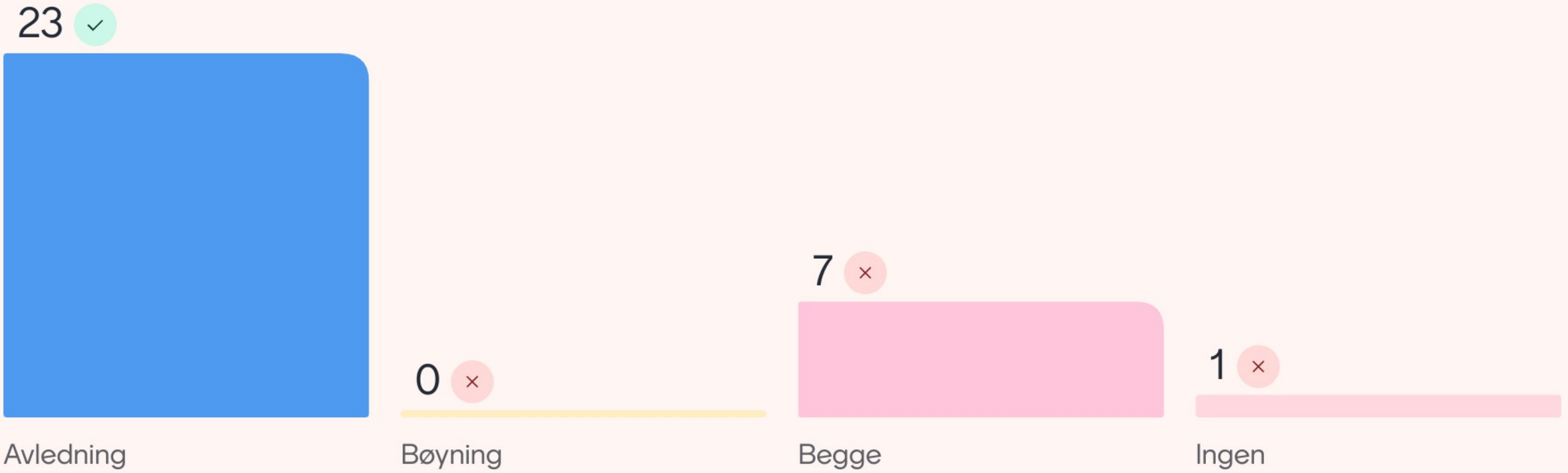
2
34

større



2
31

festlighet



4
31

Tokenisering

Hvor mange tokens og typer inneholder følgende tokeniserte tekst?

"Jul , eldre nynorsk også jol , er opprinnelig navnet på den germanske , det vil si den hedenske midtvintersfesten ."



5 32

Språklige data

Et korpus er en digitalisert samling tekster og er mye brukt innen språkteknologi.
Vi ønsker å lage et korpus for moderne norsk. Hva må vi tenke på hvis vi vil lage et korpus som er representativt for moderne norsk?
Begrunn svaret ditt med minst to ulike eksempler.

Diskuter med de ved siden av dere!
Dere har 4 minutter!

3

Ordklasser

Løsningsforslag:
Vi må tenke på å inkludere språklige data hentet fra alle former for
språklige uttrykk for at det skal være representativt.

Eksempler på dette vil være språklige data med ulike registere (tekst og tale),
sjangere (nyhetsstoff, romaner, eposter osv),
demografi (ulike kjønn, ulike aldersgrupper),
ulike dialekter osv.

1

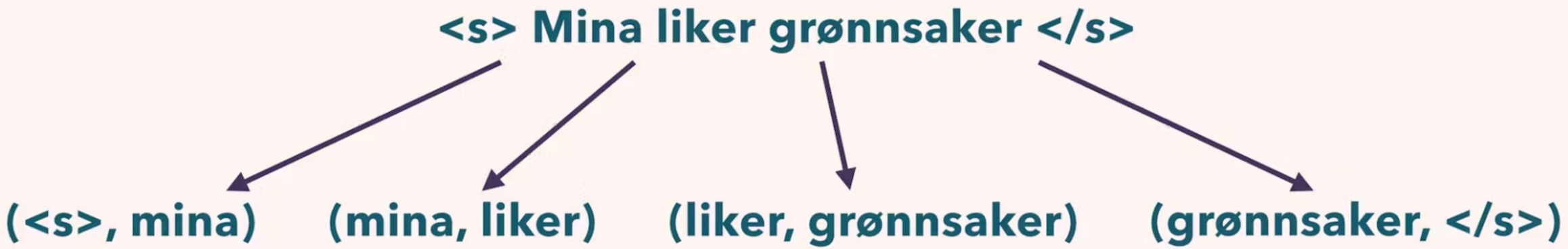
Bigram

Gitt følgende corpus:
<s> Thorvald liker grønnsaker </s>
<s> Mina liker iskrem </s>
<s> Liker broren til Mina iskrem og grønnsaker </s>

Regn ut sannsynligheten for setningen: <s> Mina liker grønnsaker </s>.

Bruk en bigram-modell.
Du kan ignorere tegnsetting og hvorvidt bokstaver er store eller små. I tillegg til den endelige sannsynligheten, skal du også vise utregningen du bruker underveis.

STEG 1: HVILKE BIGRAM TRENGER VI?



STEG 2: TELLINGER FRA KORPUSET VÅRT

<s> thorvald liker grønnsaker </s>

<s> mina liker iskrem </s>

<s> liker broren til mina iskrem og grønnsaker </s>

<s> mina liker grønnsaker </s>

3 2 3 2 3

(<s>, mina)	(mina, liker)	(liker, grønnsaker)	(grønnsaker, </s>)
1	1	1	2

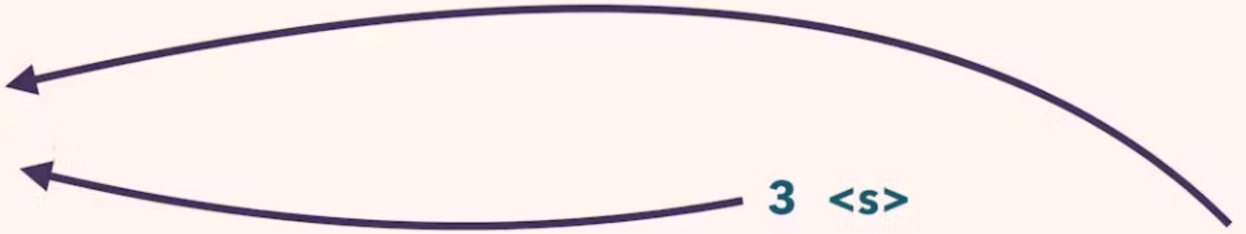
STEG 3: REGNE UT SANNSYNLIGHETER

$$P(\text{mina} \mid \text{<s>}) = \frac{1}{3}$$

$$P(\text{liker} \mid \text{mina}) = \frac{1}{2}$$

$$P(\text{grønnsaker} \mid \text{liker}) = \frac{1}{3}$$

$$P(\text{</s>} \mid \text{grønnsaker}) = \frac{2}{2}$$



3 <s>

2 mina

3 liker

2 grønnsaker

3 </s>

1 (<s>, mina)

1 (mina, liker)

1 (liker, grønnsaker)

2 (grønnsaker, </s>)

STEG 4: MULTIPLISERE SANNSYNLIGHETER

$$P(\textbf{mina} \mid \textbf{<s>}) = \frac{1}{3}$$

$$P(\textbf{liker} \mid \textbf{mina}) = \frac{1}{2}$$

$$P(\textbf{grønnsaker} \mid \textbf{liker}) = \frac{1}{3}$$

$$P(\textbf{</s>} \mid \textbf{grønnsaker}) = \frac{2}{2}$$

$$P(\textbf{<s> Mina liker grønnsaker </s>}) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} \times \frac{2}{2} = \frac{2}{36} = \frac{1}{18}$$

2

Sannsynlighet og språkmodeller

Når vi har jobbet med språkmodeller i dette emnet, har vi ganget sammen sannsynlighetene for flere n-grammer for å oppnå den samlede sannsynligheten for et helt tekststykke, for eksempel sannsynligheten for en setning.
Hva er det som kan være problematisk med en slik tilnærming?

Hva er det som kan være problematisk med en slik tilnærming?

25 ✓



For lengre tekster vil produktet bli så lite at det blir vanskelig å representere det med et normalt flyttall på datamaskinen.

0 ×

Maximum Likelihood Estimation er kun definert for setninger av en viss lengde.

2 ×

Laplace-smoothing kan resultere i udefinerte sannsynligheter som vil gi nullprodukter.

3 ×

Det er komputasjonelt tungt å regne ut produktet av alle disse sannsynlighetene.

4 30

Hvilken teoretisk antagelse fra pensum er det som uttrykkes i følgende formel, og hvorfor gjør vi den?

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(W_k|P_{k-1})$$

1

Hvilken teoretisk antagelse fra pensum er det som uttrykkes i følgende formel, og hvorfor gjør vi den?

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(W_k|P_{k-1})$$

Markov-antagelsen!

Da språket er iboende kreativt og kan skape et nærmest uendelig antall setninger, er vi nødt til å tilnærme oss sannsynligheten for en tekstsekvens fremfor å regne den ut nøyaktig. Vi kan derfor gjøre en antagelse om at sannsynligheten for et ord ikke avhenger av alle foregående ord, men kun et begrenset antall av de foregående ordene.

3

Ukjente n-gram

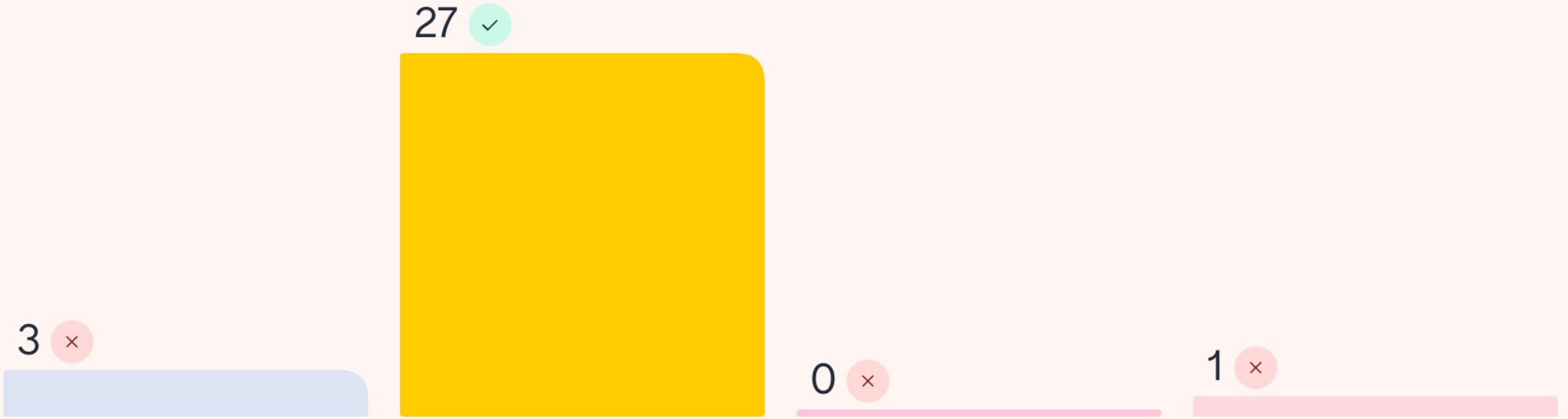
I faget har vi brukt Maximum Likelihood Estimation (MLE) for å trene n-gram modeller:

$$P(w_n|w_{n-1}) = \frac{Count(w_{n-1}, w_n)}{Count(w_{n-1})}$$

Hvordan kan vi modifisere MLE slik at modellene våre fortsatt kan brukes på tekst som inneholder nye og ukjente n-grammer?

1

Hvordan kan vi ta hensyn til usette n-gram



3 ×

I stedet for å normalisere sannsynligheten for et n-gram med frekvensen til det foregående ordet, kan vi kun normalisere med størrelsen på vokubaluret

27 ✓

Vi kan sørge for at alle mulige n-grammer har en frekvens større enn 0

0 ×

Vi kan sørge for at n-grammene er normalfordelte

1 ×

Vi kan omfordele sannsynglighetsmassen slik at høyfrekvente ord er halvparten så sannsynlige

2 31

Grammatikk

Ta utgangspunkt i følgende grammatikk og bestem om de følgende setningene er grammatiske eller ikke:

S -> NP VP

NP -> N

CP -> C S

VP -> V NP

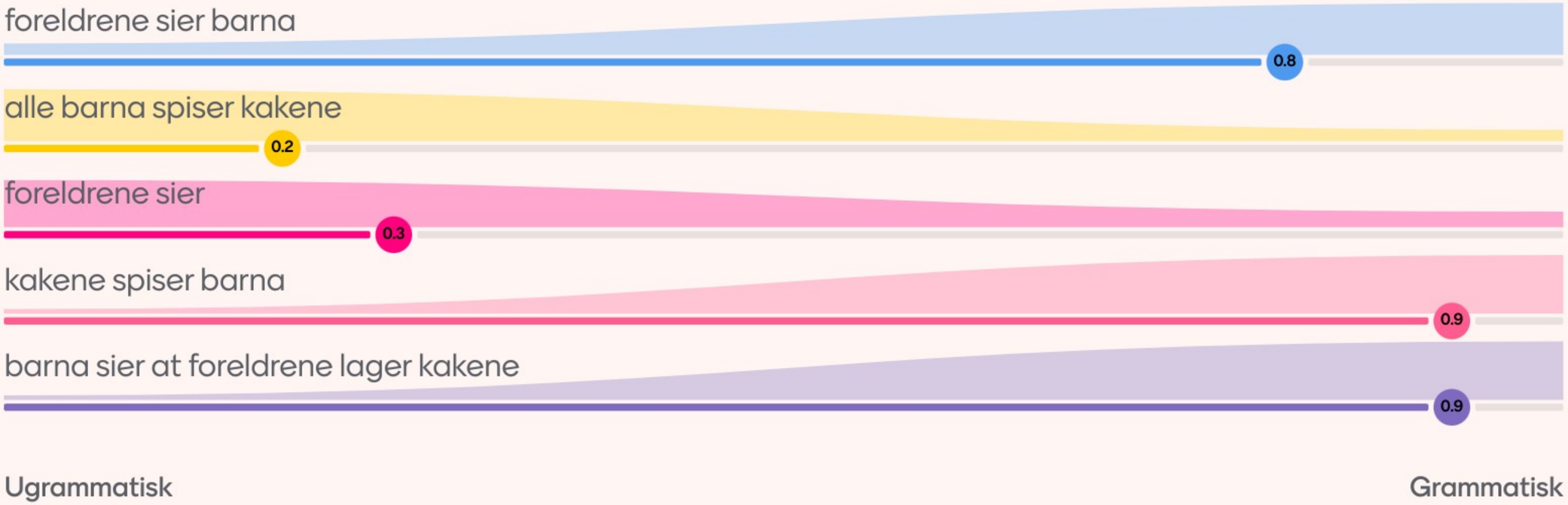
VP -> V CP

V -> spiser|sier|lager

N -> kakene|barna|kattene|foreldrene

C -> at

Hvilke setninger er grammatiske?



3
28

Diskuter i 1 minutt

Hva vil det si at en grammatikk er rekursiv?

Er grammatikken fra forrige oppgave rekursiv?

S -> NP VP

NP -> N

CP -> C S

VP -> V NP

VP -> V CP

V -> spiser|sier|lager

N -> kakene|barna|kattene|foreldrene

C -> at

2

Løsningsforslag

1. En grammatikk er rekursiv dersom den inneholder en regel som (direkte eller indirekte) introduserer en konstituent av samme type.
2. Ja, grammatikken er (indirekte) rekursiv på grunn av reglene **S -> NP VP** og **CP -> C S** som i kombinasjon introduserer en ny **S** under den øverste **S**'en.

S -> NP VP
NP -> N
CP -> C S
VP -> V NP
VP -> V CP
V -> spiser|sier|lager
N -> kakene|barna|kattene|foreldrene
C -> at

Utvidelse av grammatikk

Hvordan kan vi utvide grammatikken slik
at følgende setninger også dekkes?

1. *et barn lager en kake*
2. *en katt danser*

S -> NP VP

NP -> N

CP -> C S

VP -> V NP

VP -> V CP

V -> spiser|sier|lager

N -> kakene|barna|kattene|foreldrene

C -> at

Utvidelse av grammatikk

Hvordan kan vi utvide grammatikken slik at følgende setninger også dekkes?

1. *et barn lager en kake*
2. *en katt danser*

S -> NP VP

NP -> N | D N ←

CP -> C S

VP -> V NP | V ←

VP -> V CP

D -> en|et ←

V -> spiser|sier|lager|danser ←

N -> kakene|barna|kattene|foreldrene
|barn|kake|katt ←

C -> at

Named Entity Recognition

Vi skal analysere teksten:
St. Georg, eller Jørgen, vart fødd i Kappadokia i Vesle-Asia.
Han drepte ein drake som truga byen Silena i Libya

Nå skal du utføre "Named Entity Recognition" der vi kun har kategoriene PER og LOC.
Alle typer LOC inngår i kategorien LOC.
Tildel rett tag, eller merkelapp, på hvert ord, ved å ta i bruk BIO-tagging.

Vi prøver sammen!

St. Georg , eller Jørgen , vart fødd i Kappadokia i Vesle - Asia .
Han drepte ein drake som truga byen Silena i Libya

Fasit

St. Georg , eller Jørgen , vart fødd i Kappadokia i Vesle - Asia .

B-PER O O B-PER O O O O B-LOC O B-LOC I-LOC I-LOC O

Han drepte ein drake som truga byen Silena i Libya

O O O O O O O B-LOC O B-LOC

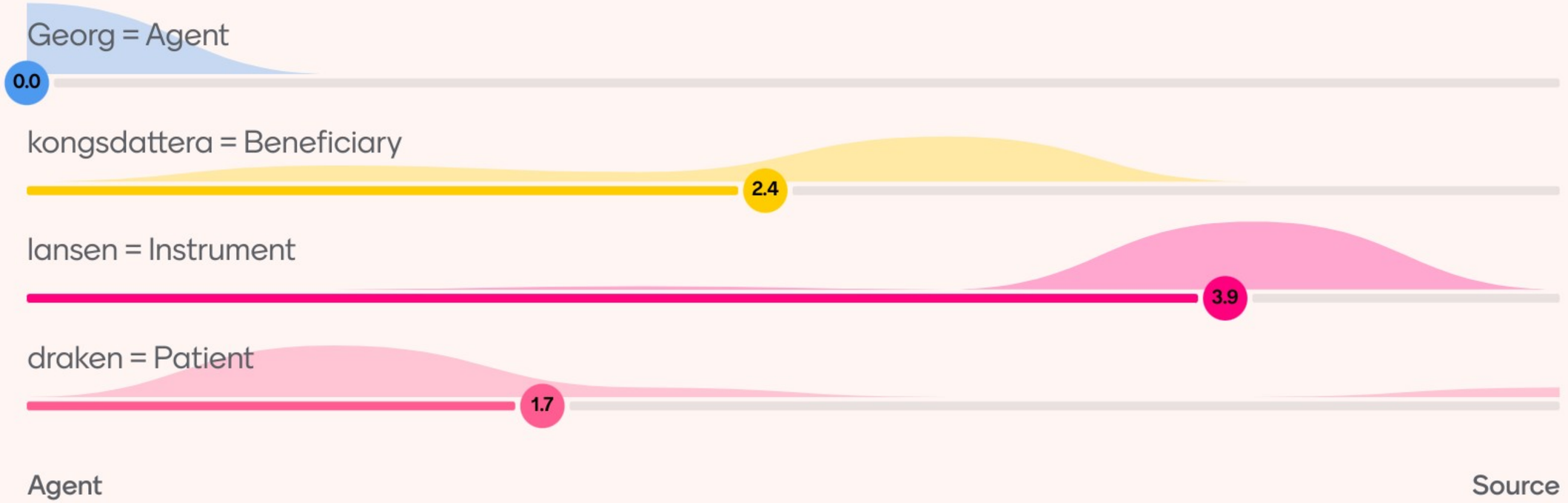
Coreference Resolution

Vi skal finne koreferanser i teksten under. Etter tokenisering er teksten delt opp som vist.
Nå skal du utføre "Coreference Resolution" med hensyn til den ene entiteten som har mer enn to referanser i teksten.
Du skal angi hvert av ordene som utgjør, eller inngår i en referanse til denne entiteten.
Disse skal ha kategorien "Cluster_01", mens de andre ordene skal ha kategorien "O".

St. Georg *, eller* ***Jørgen*** *, vart fødd i Kappadokia i Vesle - Asia .*
Han *er best kjend for at* ***han*** *drepte ein drake .*

Semantiske roller

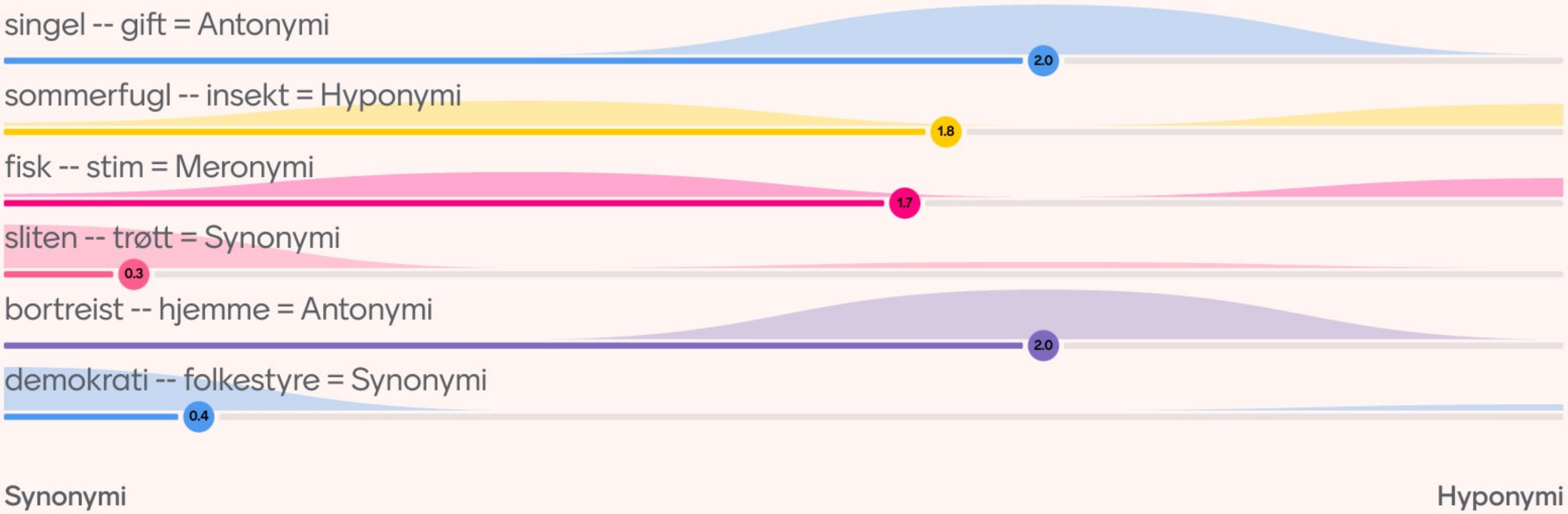
I setningen "*Georg berga kongsdottera då han stakk lansen sin gjennom draken*", hva er de semantiske rollene til:



4
22

Leksikalske relasjoner

Hvilke leksikalske relasjoner vises her?



3
16

Naive Bayes

Naive Bayes

I dette faget har vi jobbet litt med klassifisering.
Vi har som regel ønsket å finne den mest sannsynlige klassen
for et stykke tekst, gitt et sett med alternativer.
Dette kan formuleres med følgende uttrykk:

$$\hat{c} = \underset{c \in C}{argmax}\ P(c|d)$$

hvor $\hat{c}$ er den mest sannsynlige klassen av C alternativer, og d er et stykke tekst.
Vis hvordan vi kan omformulere dette uttrykket til en Naive Bayes klassifikator ved hjelp av Bayes formel.

Naive Bayes

Her må vi huske at $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

Vi kan derfor gå fra $\hat{c} = \underset{c \in C}{argmax}\ P(c|d)$

til $\hat{c} = \underset{c \in C}{argmax}\ \frac{P(d|c) * P(c)}{P(d)}$

Til slutt kan vi eventuelt stryke $P(d)$, siden denne er konstant for alle klassene.

$$\hat{c} = \underset{c \in C}{argmax}\ P(d|c) * P(c)$$

Maskinlæring

Redegjør for begrepet *veiledet læring* og hvilke komponenter som inngår i et system som bruker en slik tilnærming. Kom gjerne med eksempler fra faget.

Maskinlæring

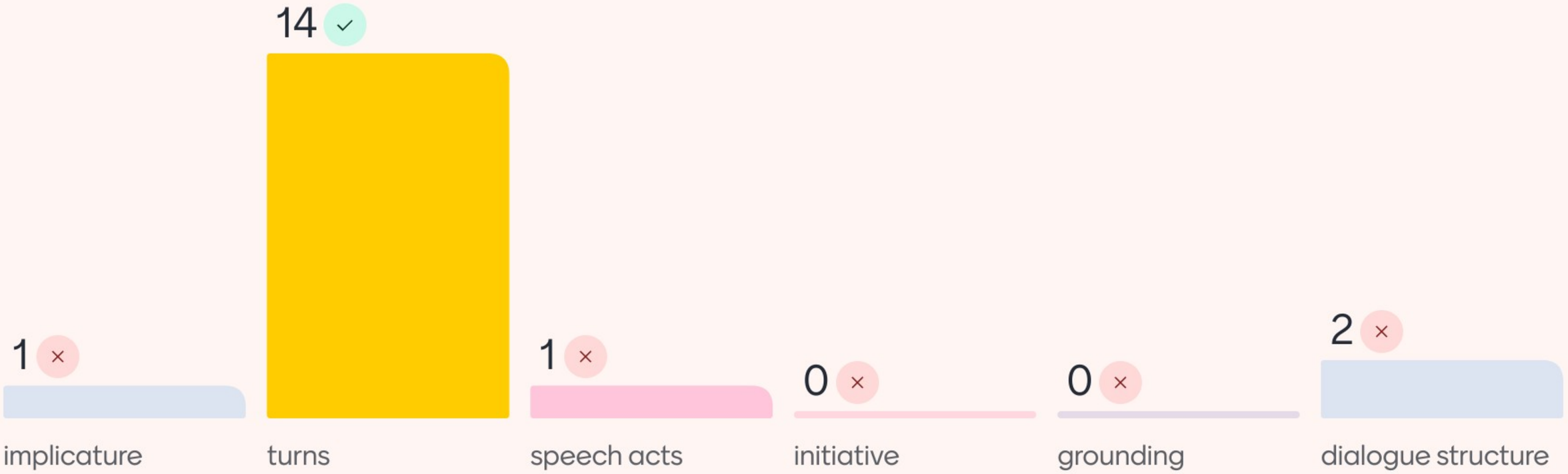
Veiledet læring = Trening av en maskinlæringsmodell ved hjelp av annoterte data.
Eksempler fra faget kan være sentimentanalyse, PoS-tagging, NER, spam-klassifisering osv.

Dialogsystemer

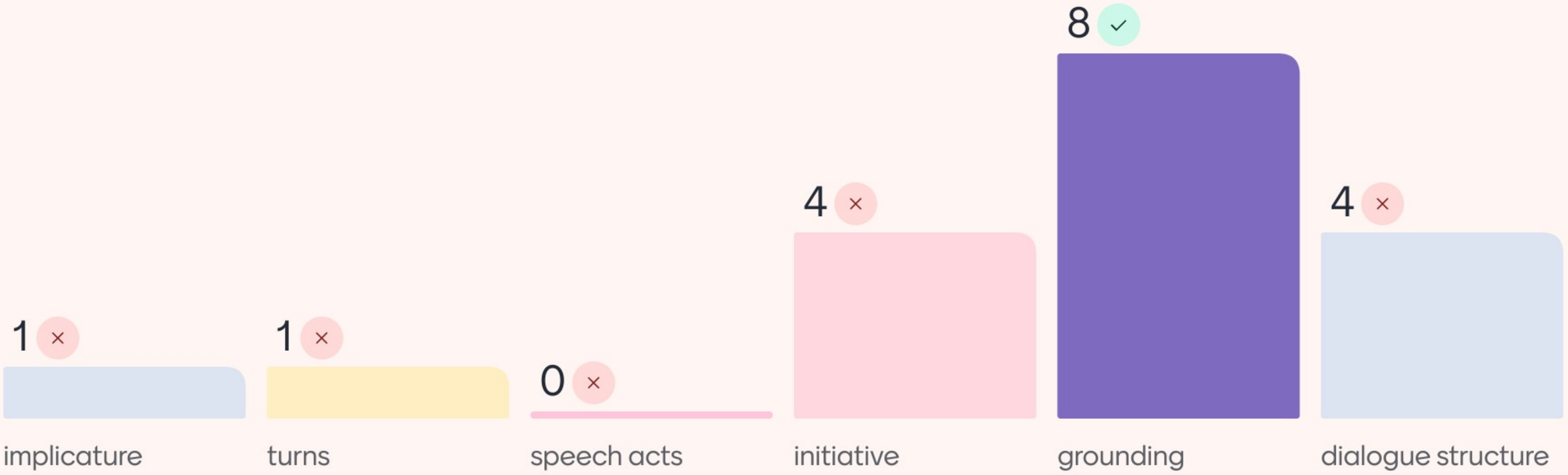
Dialog mellom mennesker har visse egenskaper som er krevende for et dialogsystem å etterligne:

turns, speech acts, grounding, dialogue structure, initiative, implicature

Endpoint detection hjelper oss å finne når en samtaleaktør er ferdig med å snakke. Det hjelper oss å skille _____ fra hverandre.

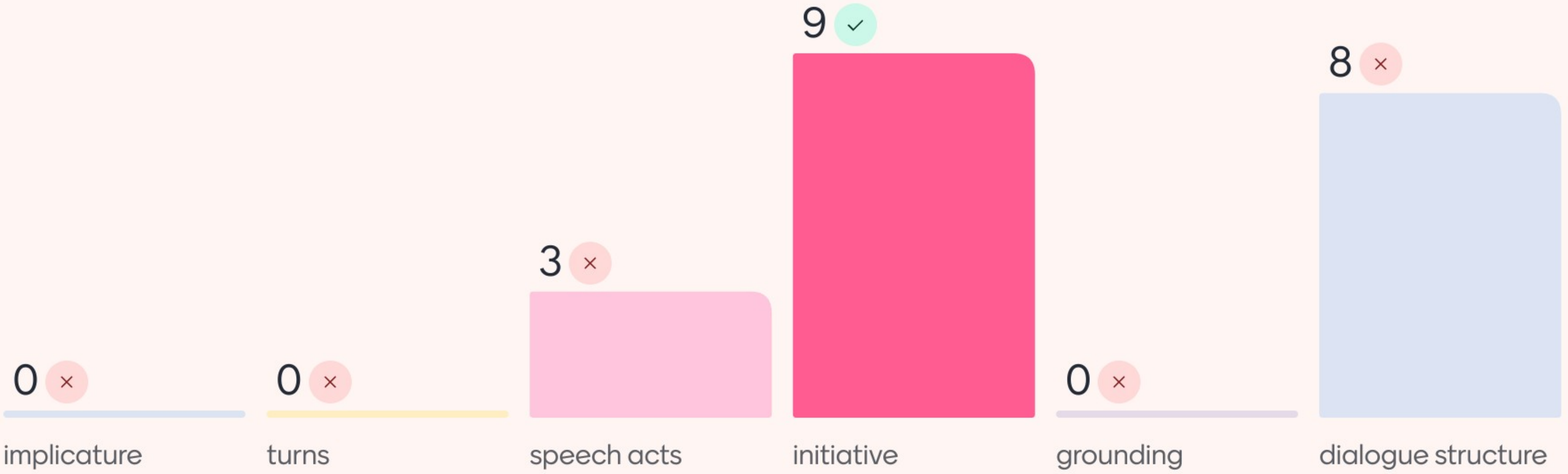


Uttrykkene "Roger", eller "Mottatt", brukes som en bekreftelse på at en beskjed er mottatt. Dette bidrar til ____ i samtalen.



1
18

Slike tilfeller der ene samtaleparten gir beskjeder som den andre tar imot er et eksempel på ubalansert _____.



2
20

Noen siste spørsmål?

Noen siste beskjeder:

- Siste gruppetimer i morgen
- Går det ikke helt bra kan man alltids prøve igjen
- Obligvinner 2024:
- nicolags 🎉🎉

Lykke til på eksamen!