TOWARD AUTOMATIC IMPROVEMENT OF LANGUAGE PRODUCED BY NON-NATIVE LANGUAGE LEARNERS

30 September 2019

Mathias Creutz and Eetu Sjöblom

WHAT WE WANT TO ACHIEVE

- Automatically correct text written by non-native speakers
- Address local errors, such as typos and incorrectly inflected forms

WHAT WE WANT TO ACHIEVE

- Automatically correct text written by non-native speakers
- Address local errors, such as typos and incorrectly inflected forms
- Also address larger issues in order to make the text more idiomatic and make it sound more natural
  - Related to paraphrasing (similar semantic content, but different surface structure)
  - For example, She blindsided me. ⇔ She took me by surprise.

WHAT WE WANT TO ACHIEVE

- Automatically correct text written by non-native speakers
- Address local errors, such as typos and incorrectly inflected forms
- Also address larger issues in order to make the text more idiomatic and make it sound more natural
  - Related to paraphrasing (similar semantic content, but different surface structure)
  - For example, She blindsided me. ⇔ She took me by surprise.
- Could evolve into an automatic tool for language learners / users

TEST DATA

yki-korpus.jyu.fi/?lang=en

Search

Yki-Korpus | Login | Choose language | Help

Welcome to YKI-Corpus!

The corpus has been compiled from the examinations of The National Certificates of Language Proficiency. It is intended for research purposes.

The National Certificates of Language Proficiency is a language testing system for adults and is not connected to any curriculum or syllabus. Examinations can be taken in nine languages: English, Finnish, German, Italian, North Sami, Russian, Spanish, and Swedish. There are three test levels: Basic, Intermediate and Advanced offering six levels of proficiency (1-2, 3-4, 5-6). The corpus contains data from all nine languages and levels.

The corpus contains both quantitative and qualitative data. Quantitative data provide corpus users with assessments of the four skills of reading, writing, listening, and speaking (structures and vocabulary prior 2012) as well as background information given by the candidates during the examinations. Because filling in the background information questionnaire is voluntary, this data are not available on every candidate.

Qualitative data contain oral and written performances, i.e. candidates' responses to the tests tasks. There are three written performances and one oral performance from each candidate.

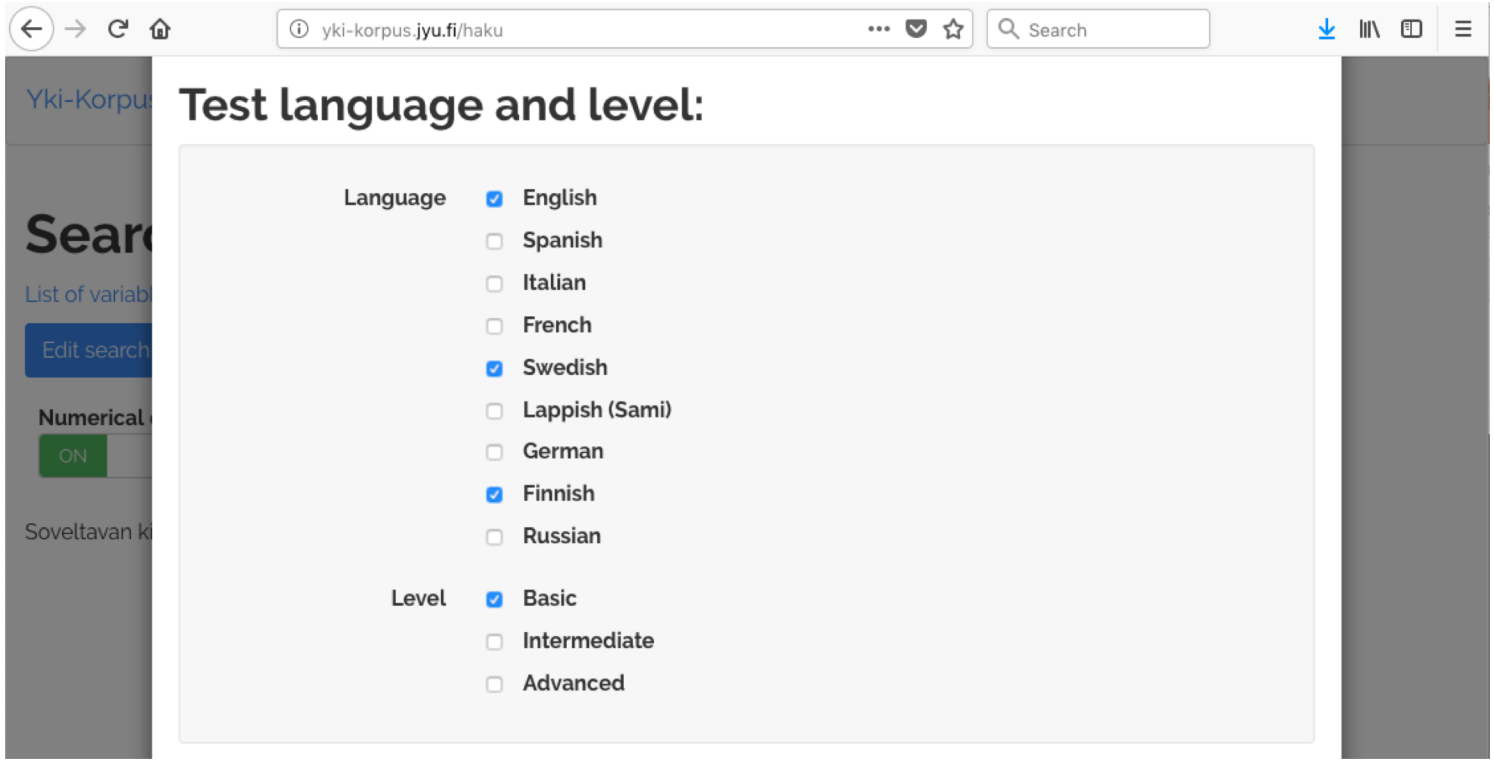
The data are connected with candidates' id numbers, which makes it possible to search between different data. The data are not necessarily equal in size.

The corpus is dynamic in nature, new data are added after each test round.

More information about the National Certificates of Language Proficiency at www.oph.fi and at www.solki.jyu.fi

Soveltavan kielentutkimuksen keskus © 2019

TEST DATA


yki-korpus.jyu.fi/haku
Search
Yki-Korpus
Sear
List of variabl
Edit search
Numerical
ON
Soveltavan ki
Test language and level:
Language
English
Spanish
Italian
French
Swedish
Lappish (Sami)
German
Finnish
Russian
Level
Basic
Intermediate
Advanced

ENGLISH EXAMPLES

Text # 1

- Text type: Informal letter
- Task type : Guided
- Headline: Kutsu

Hello Jari

I have a party in my home
next weekend. Come Saturday
evening at 18.00. Can you bring
some food and couple beer
with you. Here comes about
eight my friends. you can sleep
here at night.

I hope to see you.
Matti

Text # 2

- Text type: Formal message
- Task type : Guided
- Headline: Viesti

My name is Maija Solki. I have writing,
because I don't speak English well, but
I would like learn much more.
I had learn english about six year.

Text # 3

- Text type: Opinion
- Task type : Guided
- Headline: Mielipide

That's very difficult problem for you and
perhaps to all people in your flat.
The first you must to talk to these
young people for this problem. Ask them to stop like this
If you don't get help for it then you
must to tell for your house/flat "isännöitsijä"
If it isn't help then you will ask for help
to polishman :-)
Good luck!

FINNISH AND SWEDISH EXAMPLES

Text # 2

- Text type: Informal message
- Task type : Guided
- Headline: Viesti

Eilen kun tulin kotiin koulusta minä

löytin uksi lauku, Solkipolku kadulla.
Se on ihan iso, se on tossi pehmeä,
mutta punainen, minä en tutka
punaisea vatet eikä laukot tai kengät.
Sisällä se oli henkilotieto. Se on
Maijan lauku, minä tietä kuka hän
on koska hän on opiskelija säma koulu
kun minä.
Varma tannan minä voi anta hanelle.
Tai viime viiko koulussa
Toivotavasti hän ei tarvitsee kaiki tavarat
tämä vikkonloppu.

Text # 1

- Text type: Informal letter
- Task type : Guided
- Headline: Kort

Hei Anne!

Hur mår du? Jag mår bra!

Jag tycker om min nya bostad.

Men det varit första lite svårt här att

finna nya kompisar.

Men nu jag giller om vara

här! Staden är stor

och mycket vänligt. Men

vi har också naturen i när.

Jag spelar också cello här!

I en orkestra, det är jätte

viktigt för mej! På fritiden

jag spela med Sanna min hund

på stranden i när.

Dina Maija

TRAINING DATA

- OpenSubtitles corpus (Lison and Tiedemann, 2016)

- We have used the parallel subcorpora (bitexts):
  - Finnish – English: *23 million sentence pairs*
  - Finnish – Swedish: *12 million sentence pairs*
  - Swedish – English: *15 million sentence pairs*

TRAINING DATA

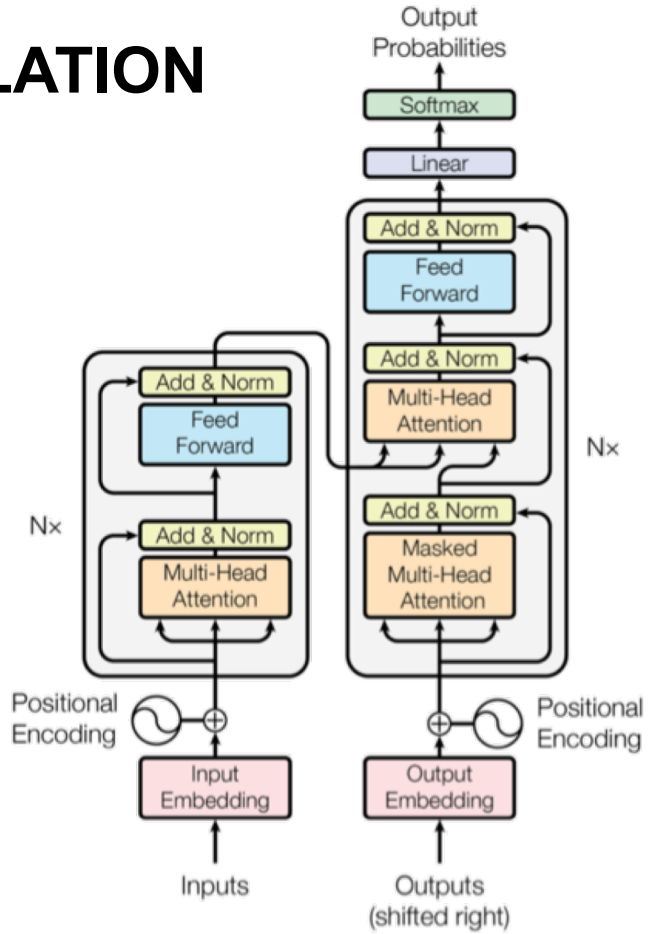
- OpenSubtitles corpus (Lison and Tiedemann, 2016)

- We have used the parallel subcorpora (bitexts):
  - Finnish – English: *23 million sentence pairs*
  - Finnish – Swedish: *12 million sentence pairs*
  - Swedish – English: *15 million sentence pairs*

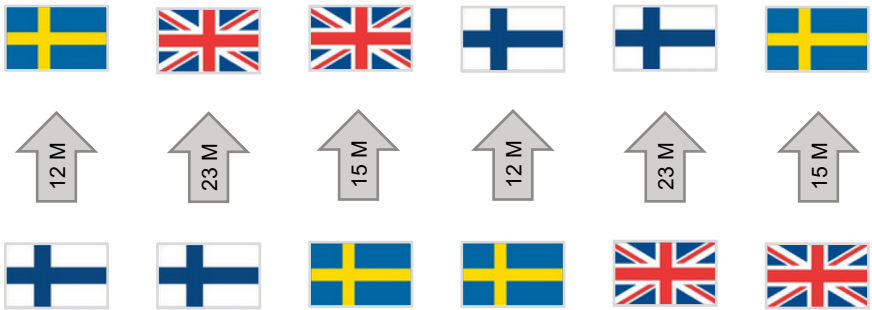
You get cross so easily . (en) ⇔ Hermostut niin pienestä . (fi)
Jag måste prata med henne just nu . (sv) ⇔ I have to talk to her right now . (en)
– Tiedän että opit. (fi) ⇔ – Det vet jag att du vill . (sv)

NEURAL MACHINE TRANSLATION

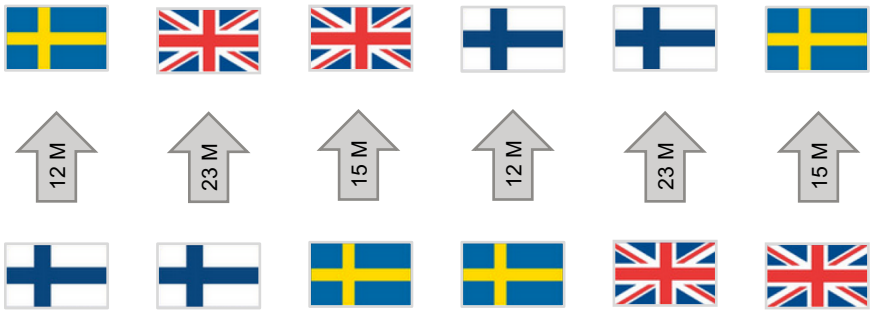
- Transformer model (Vaswani et al., 2017)
- OpenNMT-py (Klein et al., 2017)
- Character-based models
- Target-language flag (Johnson et al., 2017)



TRAINING A MODEL: 1. BASELINE

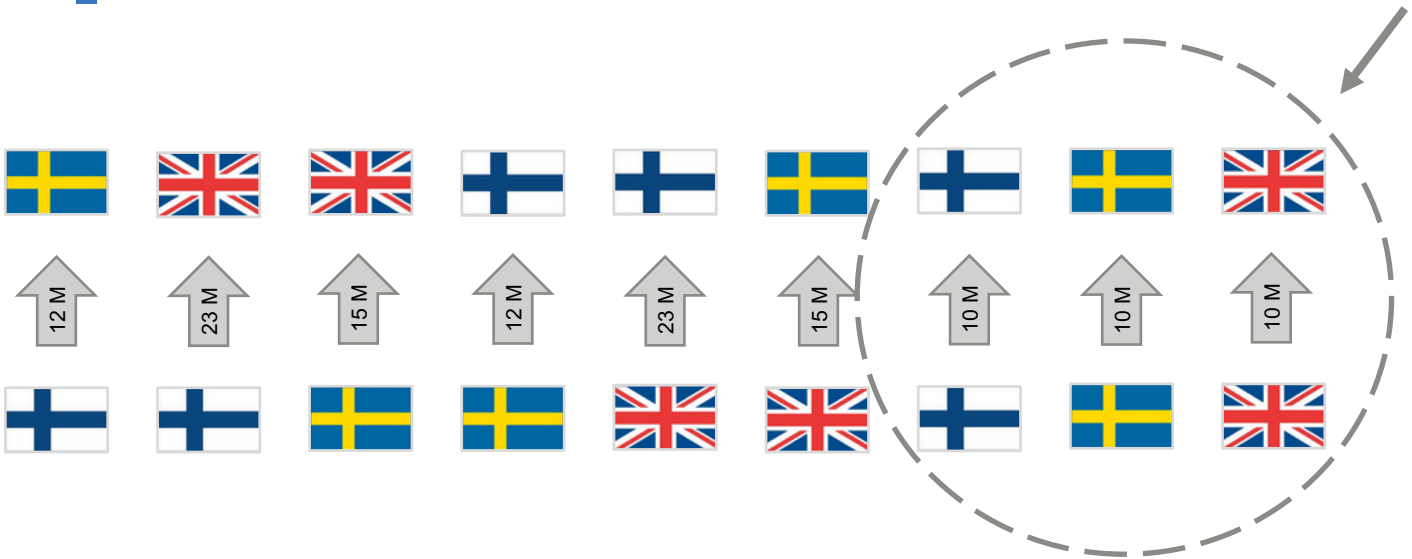


TRAINING A MODEL: 1. BASELINE

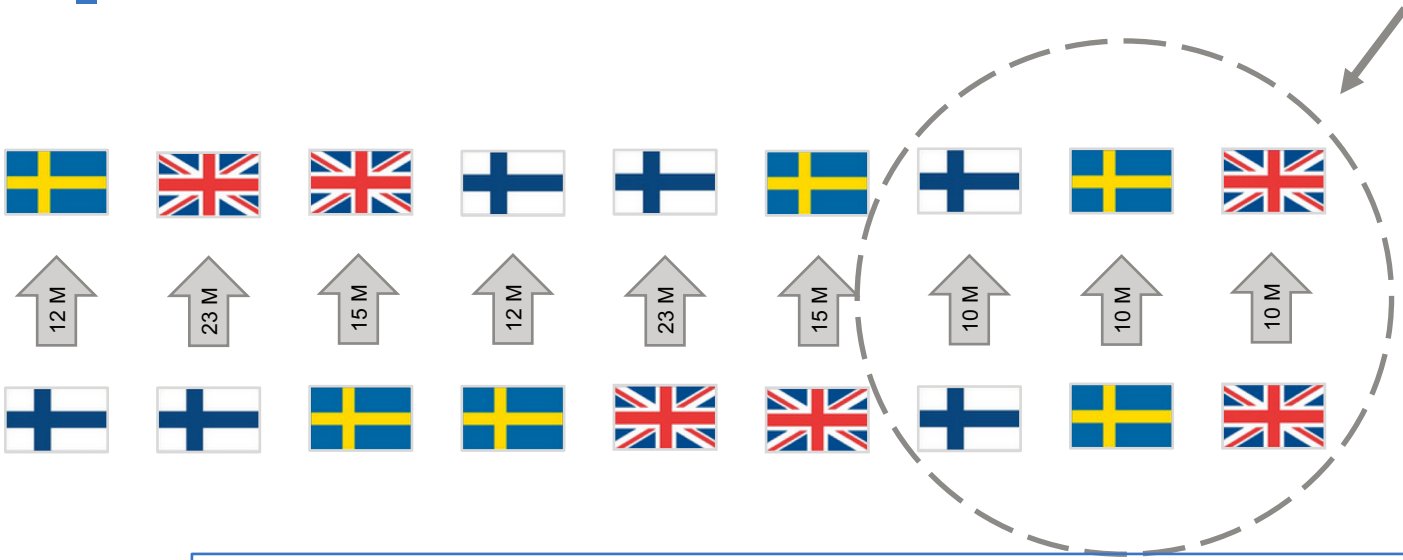


[[fi]] Y o u | g e t | c r o s s | s o | e a s i l y | . ⇨ H e r m o s t u t | p i e n e s t ä | .
[[en]] J a g | m å s t e | p r a t a | m e d | h e n n e | j u s t | n u | . ⇨
I | h a v e | t o | t a l k | t o | h e r | r i g h t | n o w | .
[[sv]] – | T i e d ä n | e t t ä | o p i t | . ⇨ – | D e t | v e t | j a g | a t t | d u | v i l l | .

TRAINING A MODEL: 2. BASELINE + CLONES



TRAINING A MODEL: 2. BASELINE + CLONES

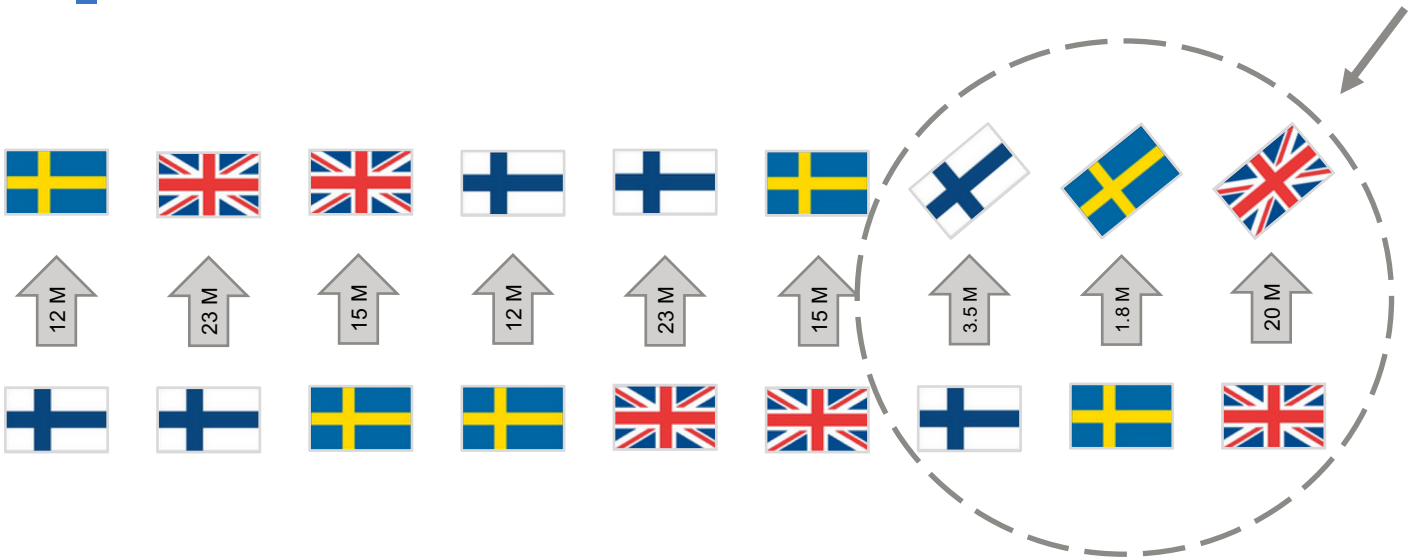


[[en]] Y o u | g e t | c r o s s | s o | e a s i l y | . ⇨ Y o u | g e t | c r o s s | s o | e a s i l y | .

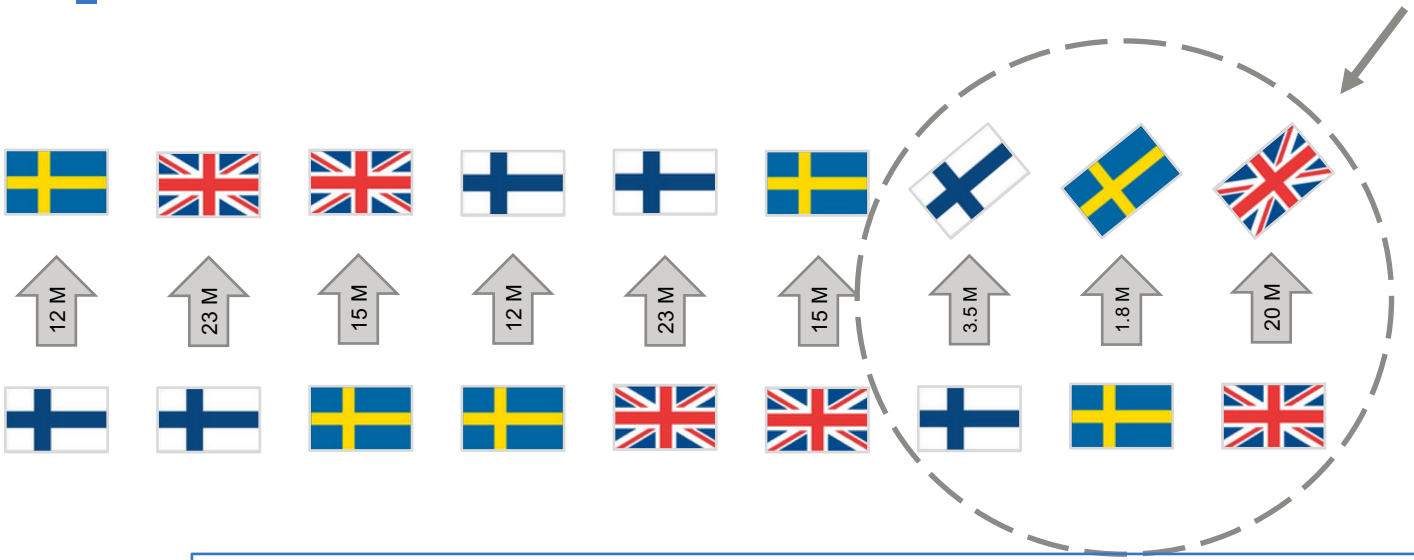
[[sv]] J a g | m å s t e | p r a t a | m e d | h e n n e | j u s t | n u | . ⇨ J a g | m å s t e | p r a t a | m e d | h e n n e | j u s t | n u | .

[[fi]] – | T i e d ä n | e t t ä | o p i t | . ⇨ – | T i e d ä n | e t t ä | o p i t | .

TRAINING A MODEL: 3. BASELINE + OPUSPARCUS



TRAINING A MODEL: 3. BASELINE + OPUSPARCUS

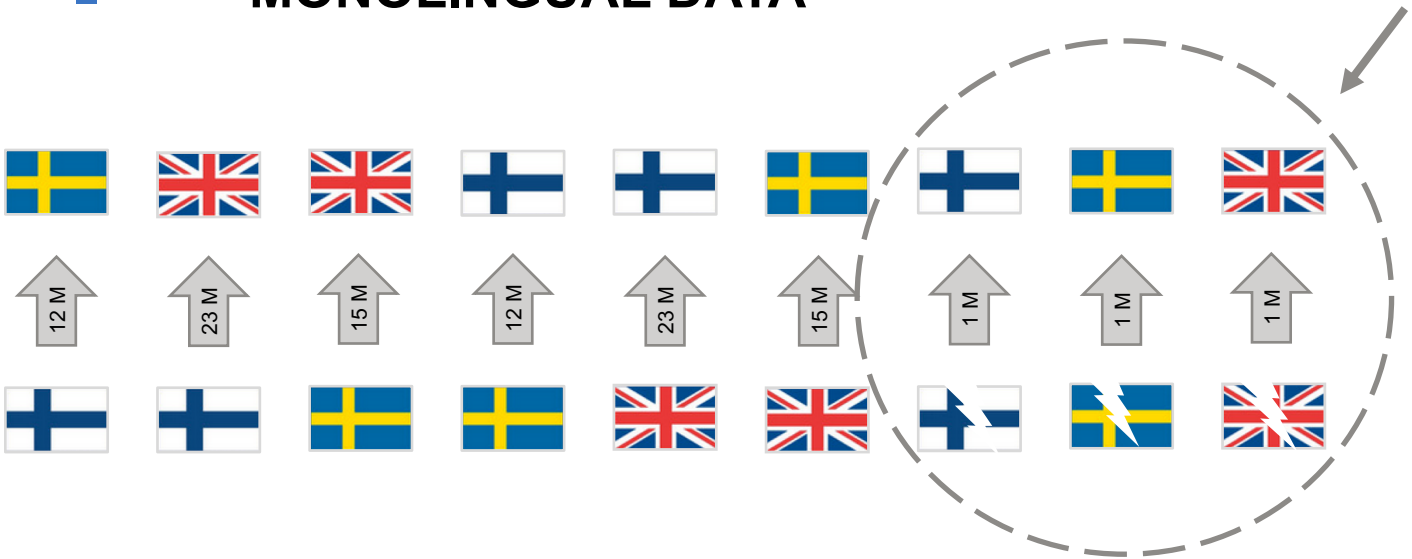


[[en]] H o w | m a n y | c h i l d r e n | d o | y o u | h a v e | ? ⇨ H o w | m a n y | k i d s | h a v e | y o u | g o t | ?

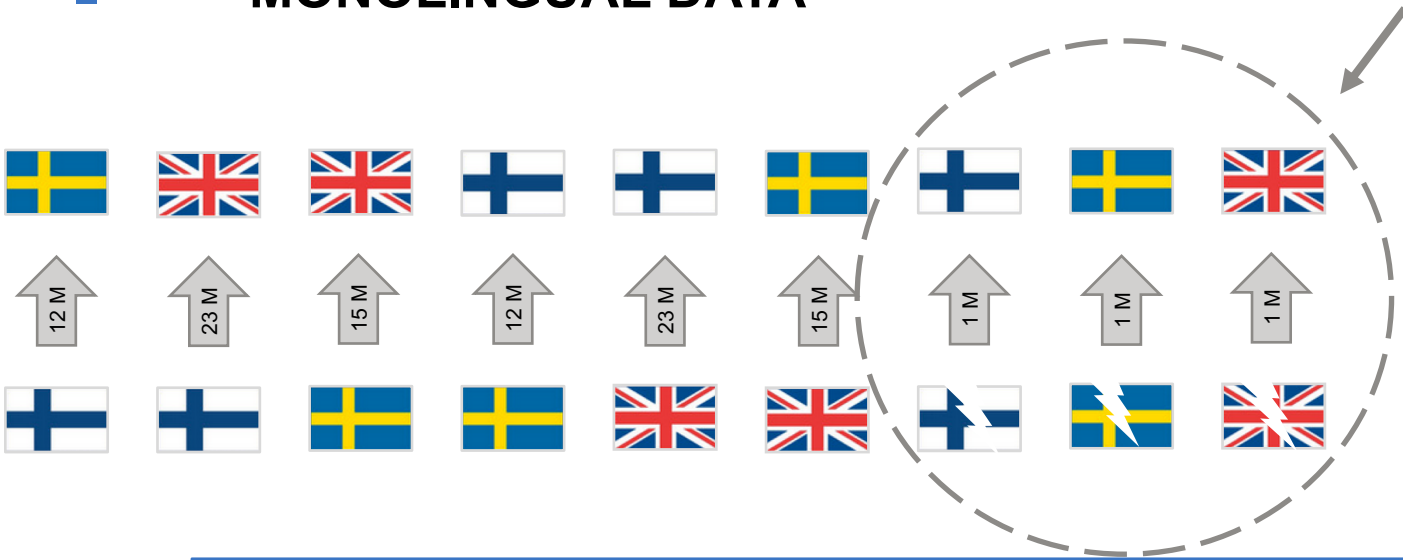
[[sv]] P a p p a | d ö d a r | m ä n n i s k o r | . ⇨ P a p p a | m ö r d a r | f o l k | .

[[fi]] E i | o l e | k o l i k o i t a | . ⇨ M i n u l l a | e i | o l e | p i k k u r a h a a | .

TRAINING A MODEL: 4. BASELINE + ERROR-AUGMENTED MONOLINGUAL DATA



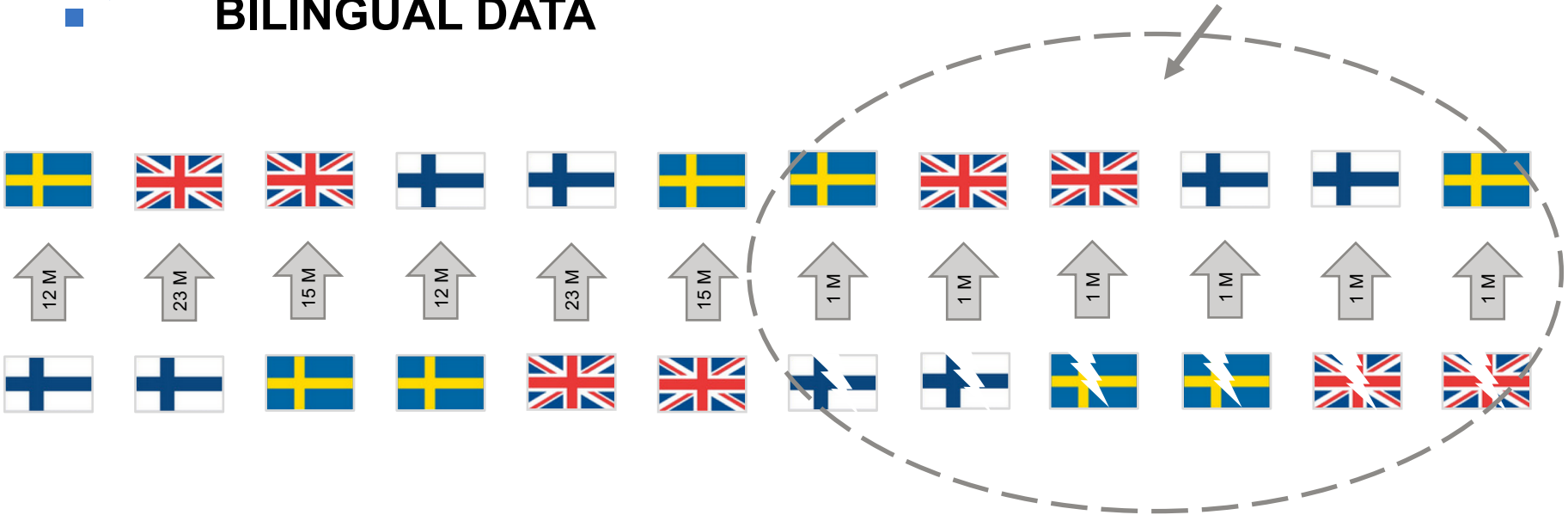
TRAINING A MODEL: 4. BASELINE + ERROR-AUGMENTED MONOLINGUAL DATA



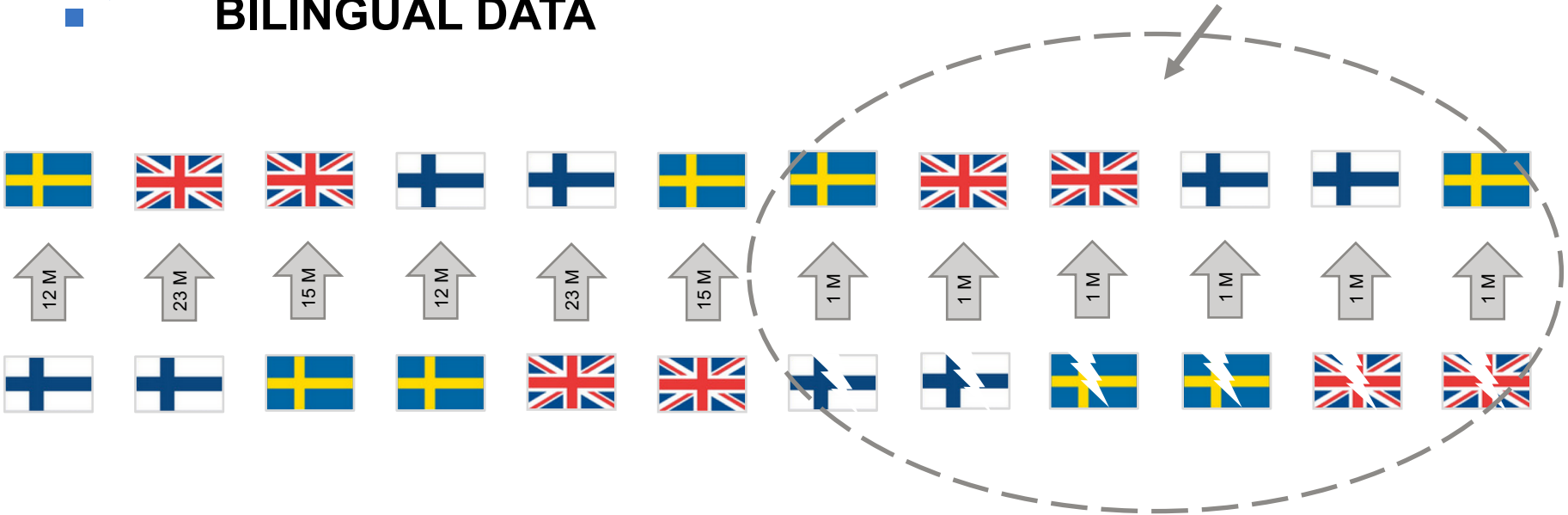
[[fi]] **A e** | t a s k u s s a | n ä k ö j ä ä n | **v o i t t s a** | **t i k a r r i n** | s a a p p a a s s a | . ⇨
A s e | t a s k u s s a | n ä k ö j ä ä n | v o i t t a a | t i k a r i n | s a a p p a a s s a | .

[[en]] A f t e r | s h e | **a t t a c k s** | y o u | , | p e r h a p s | y o u | **h a d** | **s e e** | h e r | ? ⇨
A f t e r | s h e | a t t a c k e d | y o u | , | p e r h a p s | y o u | h a v e | s e e n | h e r | ?

TRAINING A MODEL: 5. BASELINE + ERROR-AUGMENTED BILINGUAL DATA



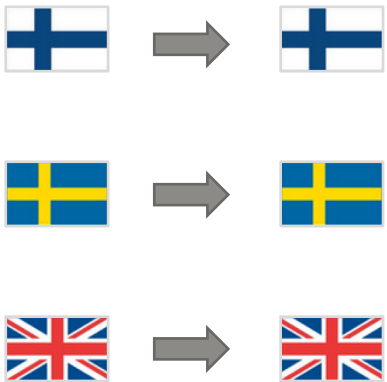
TRAINING A MODEL: 5. BASELINE + ERROR-AUGMENTED BILINGUAL DATA



[[fi]] I | h a d | t o | **g o t** | t h e | b i g g e r | **o n e** | **’ s** | . ⇨ P i t i | s a a d a | i s o m p i | .

TESTING TRANSLATION MODELS ON YKI CORPUS DATA

i) Monolingual translation (”zero shot”)



MONOLINGUAL TRANSLATION: RESULTS

1. Baseline	
2. Baseline + Clones	
3. Baseline + Opusparcus	
4. Baseline + Error-Augmented Monolingual Data	
5. Baseline + Error-Augmented Bilingual Data	

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline	
2. Baseline +	
3. Baseline +	
4. Baseline +	
5. Baseline +	

MIXED LANGUAGE INFLUENCE:

Thank you for the gift and that you remembered my birthday . ⇨

Thank you for the **poison** and you remembered my **first day** .

I think everybody is laughing to me . ⇨ I think everybody 's **screaming** for me .

Olisi ihana käydä kursilla myös kesällä . ⇨ Olisi kiva käydä kurssilla **summer**illa .

Minä opin soittaa kitaraa itsestäni ja kitarani maksoi vain 100 € . ⇨

Minä **lärvin** soittamaan kitaraa minulle , ja kitara **kosti** vain 100 .

Jag har en skönt lägenhet . ⇨ Jag har en bra **apa** .

Det är ett bra dator och svart screen . ⇨ Det är en bra **kompis** och **black** skärm .

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcu

OUTPUT = INPUT (no error correction)

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + E

5. Baseline + E

SUMMARIZATION:

But everybody should accept other people and try to not disturb them .

⇨ Everyone should accept other people .

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline +

3. Baseline +

TOO ERROR TOLERANT (local changes mostly):

If you faring near to Helsinki you are wel came to my flat in Itäkeskus . ⇨

If you fare near to Helsinki you are well came to my flat in Itäkeskus .

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

MONOLINGUAL TRANSLATION: RESULTS

1. Baseline

2. Baseline + Clones

3. Baseline + C

4. Baseline + E

MIXED LANGUAGE INFLUENCE (again):

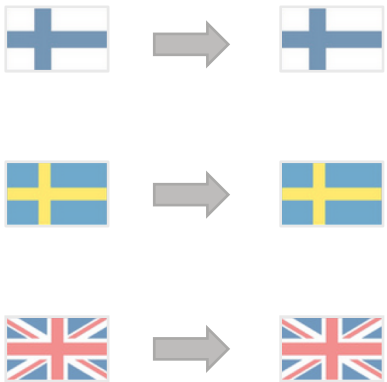
I'd like only some money for my travel box . ⇨

I just want some **penguins** for my **restaurant** .

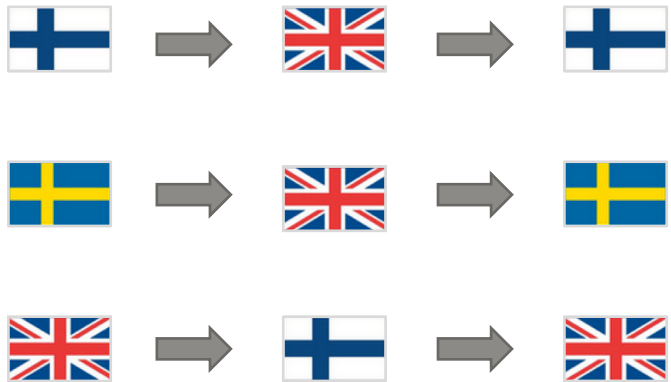
5. Baseline + Error-Augmented Bilingual Data

TESTING TRANSLATION MODELS ON YKI CORPUS DATA

i) Monolingual translation (”zero shot”)



ii) Round-trip translation (”pivoting”)



EXAMPLES OF SUCCESSFUL ROUND-TRIP TRANSLATION

- Other chance is not . ⇨ Toinen mahdollisuuus ei ole. ⇨ There ’s no other chance .
- Depence if you take with new tires or with old one ’s . ⇨ Riippuu , otatko uusia renkaita tai vanhan renkaita . ⇨ Depends if you take new tires or old tires .
- Minä löydäin pientä koiraa kadulla . ⇨ I found a little dog on the street . ⇨ Löysin pienen koiran kadulta .
- Jag måste hjälpa min systorna på en school - projekten . ⇨ I have to help my sisters on a school project . ⇨ Jag måste hjälpa mina systrar på ett skolprojekt .

UNDESIRABLE TRANSLATION OF PROPER NAMES

- I ’ll like ti meet you for my place in Rantakylä. ⇨ Tavataan ensi viikonloppuna Rantakylässä . ⇨ Meet me next weekend in **Beach Village** .
- The celebration is in Savitaipale on Olkkolan kartano . ⇨ Juhlat ovat Savitaipalessa Olkkolan Kartanossa . ⇨ The party is in Savitaipale in **Olcola** Mansion .

ROUND-TRIP TRANSLATION: WHICH MODEL PERFORMS BEST?

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

5. Baseline + Error-Augmented Bilingual Data

ROUND-TRIP TRANSLATION: WHICH MODEL PERFORMS BEST?

1. Baseline

2. Baseline + Clones

3. Baseline + Opusparcus

4. Baseline + Error-Augmented Monolingual Data

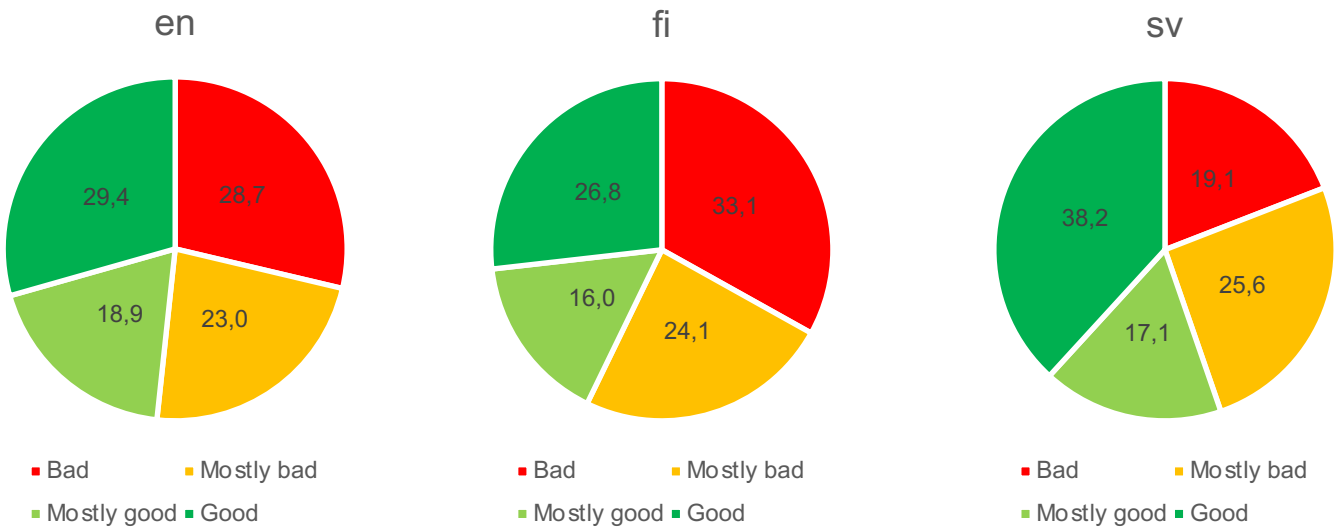
5. Baseline + Error-Augmented Bilingual Data

MANUAL ANNOTATION

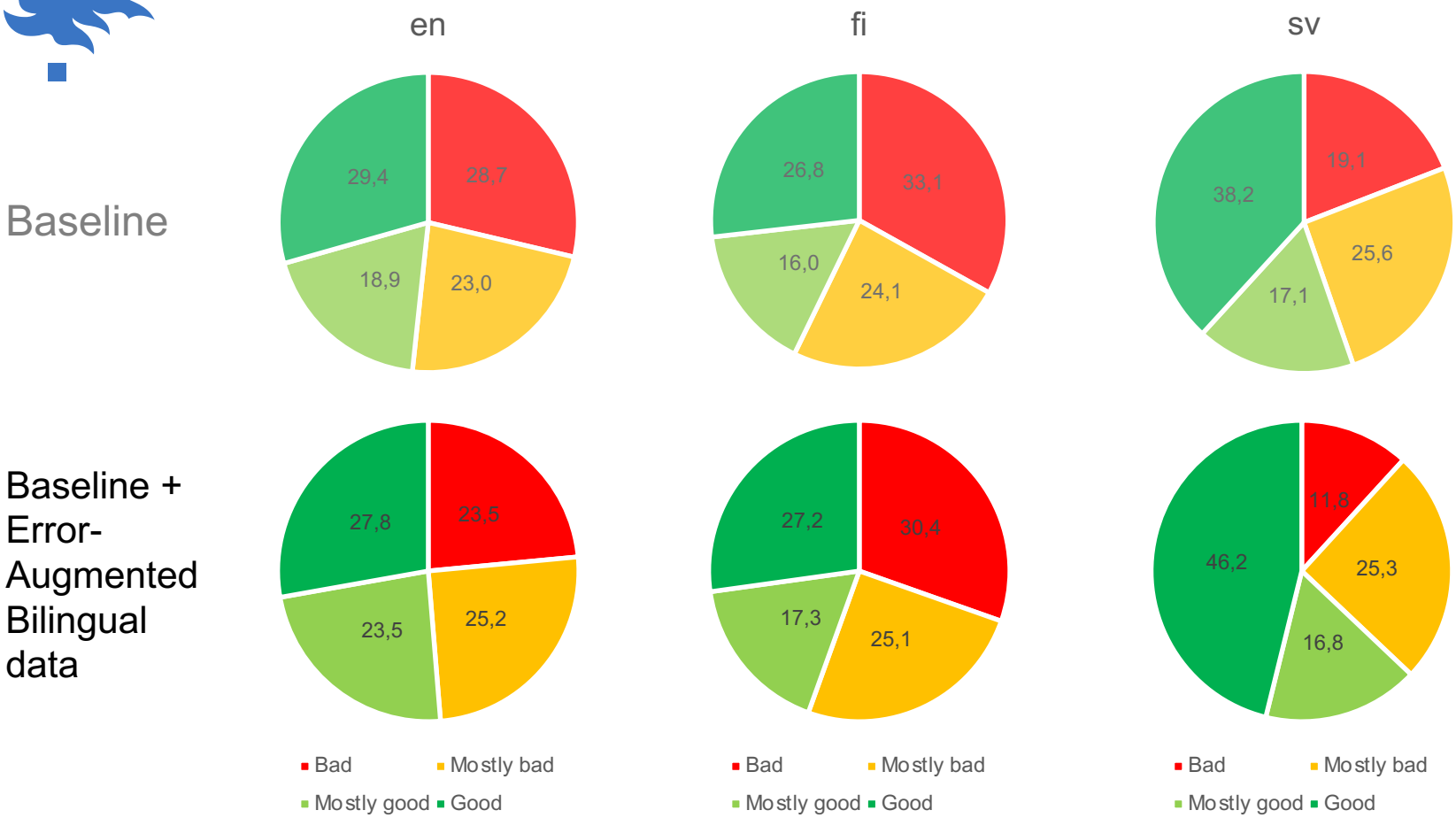
- Four annotation categories:
  1. Bad
  2. Mostly bad
  3. Mostly good
  4. Good
- Correctness and semantic adequacy in the same score
- Moderate inter-annotator agreement according to Cohen's Kappa

RESULTS OF MANUAL ANNOTATION

Baseline



RESULTS OF MANUAL ANNOTATION



WHAT NEXT?

- Get rid of round-trip translation
- Automatic evaluation
- New data sets?
- Extract ”structurally interesting” sentences to correct?

THANK YOU!