the duolingo CEFR checker:

a multilingual tool for adapting learning content

Bill McDowell

December 2021

disclaimer

- **Outdated** — All results are from 3 years ago
  - My memory is fuzzy about many details
  - There is relevant recent work on large language models and contextual embeddings

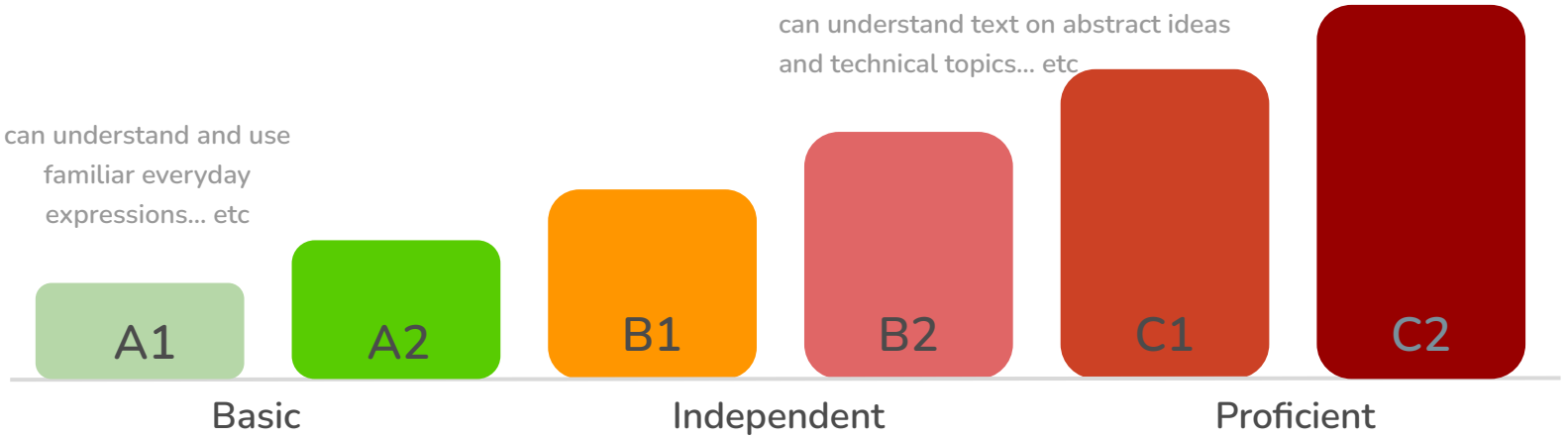
disclaimer

- **Outdated** — All results are from 3 years ago
  - My memory is fuzzy about many details
  - There is relevant recent work on large language models and contextual embeddings
- **Practical and a little messy** — Aimed to get things working, not to exhaustively test hypotheses
  - Some obvious experiments are missing
  - Some relevant result breakdowns are scattered across old spreadsheets, so I only report those which were readily available

disclaimer

- **Outdated** — All results are from 3 years ago
  - My memory is fuzzy about many details
  - There is relevant recent work on large language models and contextual embeddings
- **Practical and a little messy** — Aimed to get things working, not to exhaustively test hypotheses
  - Some obvious experiments are missing
  - Some relevant result breakdowns are scattered across old spreadsheets, so I only report those which were readily available
- **Unreviewed and unpublished** — Don’t take anything here too seriously

CEFR

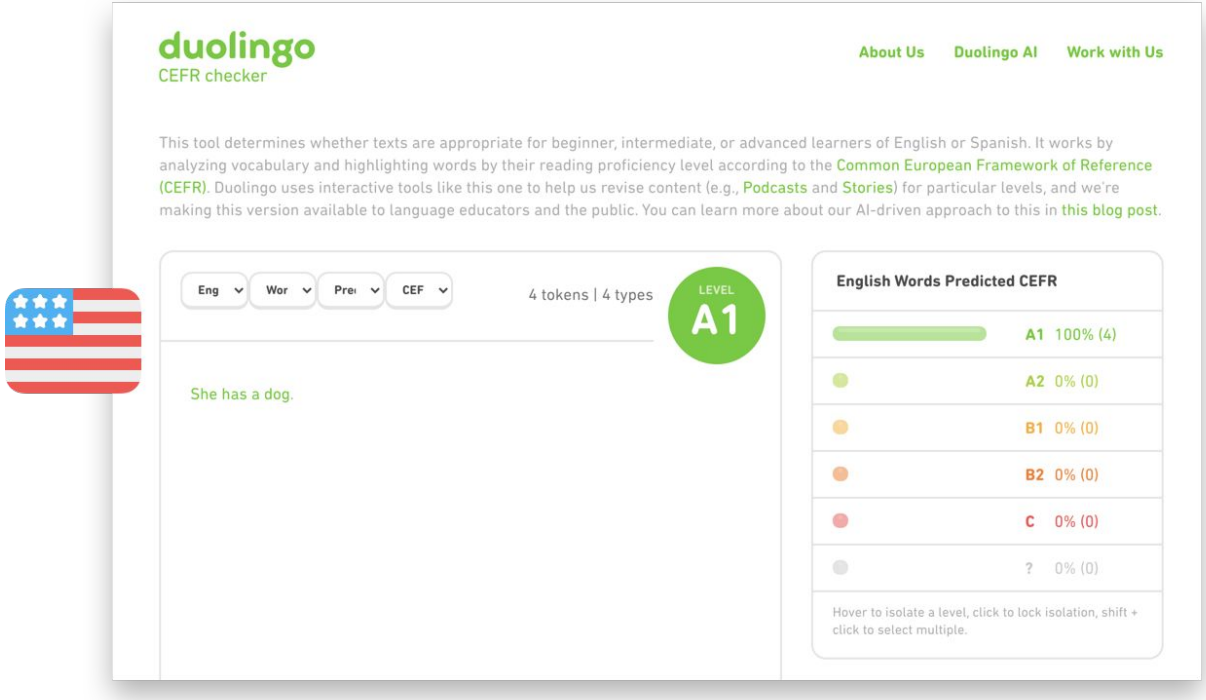

can understand and use familiar everyday expressions... etc
can understand text on abstract ideas and technical topics... etc
A1
A2
B1
B2
C1
C2
Basic
Independent
Proficient

CEFR checker

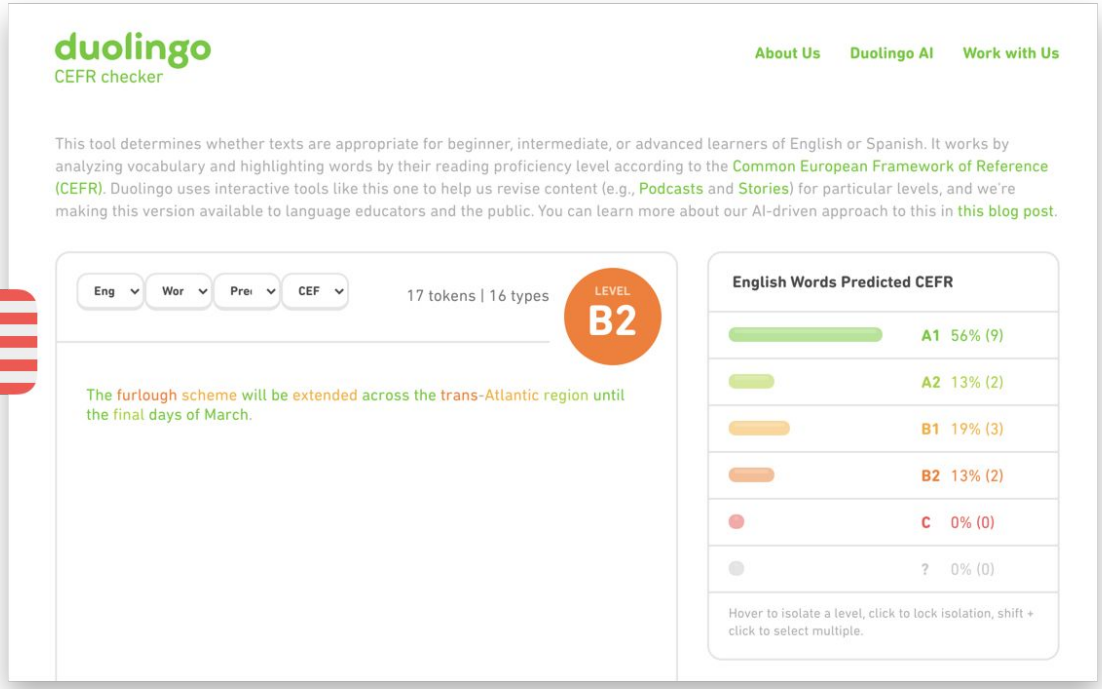
- **Check appropriate proficiency levels for content** — Automatically determine whether text is appropriate for language learners at various CEFR levels

- **Multilingual** — Work across multiple languages to adapt content for many Duolingo courses
    - English, French, Spanish, German, Italian, Portuguese

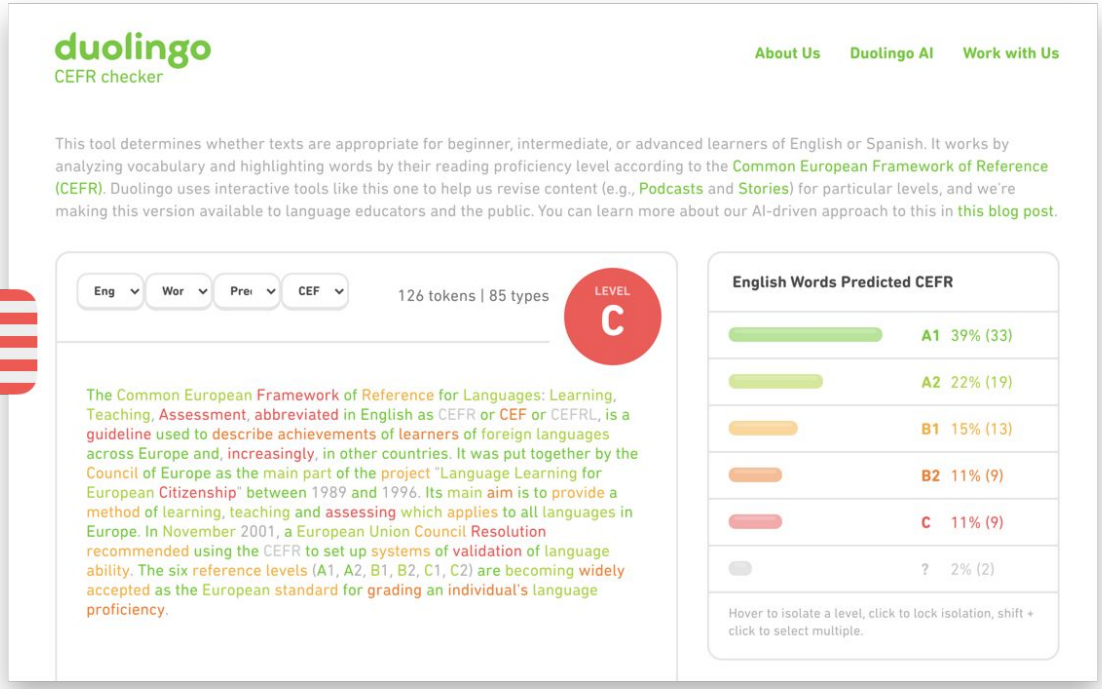
CEFR checker



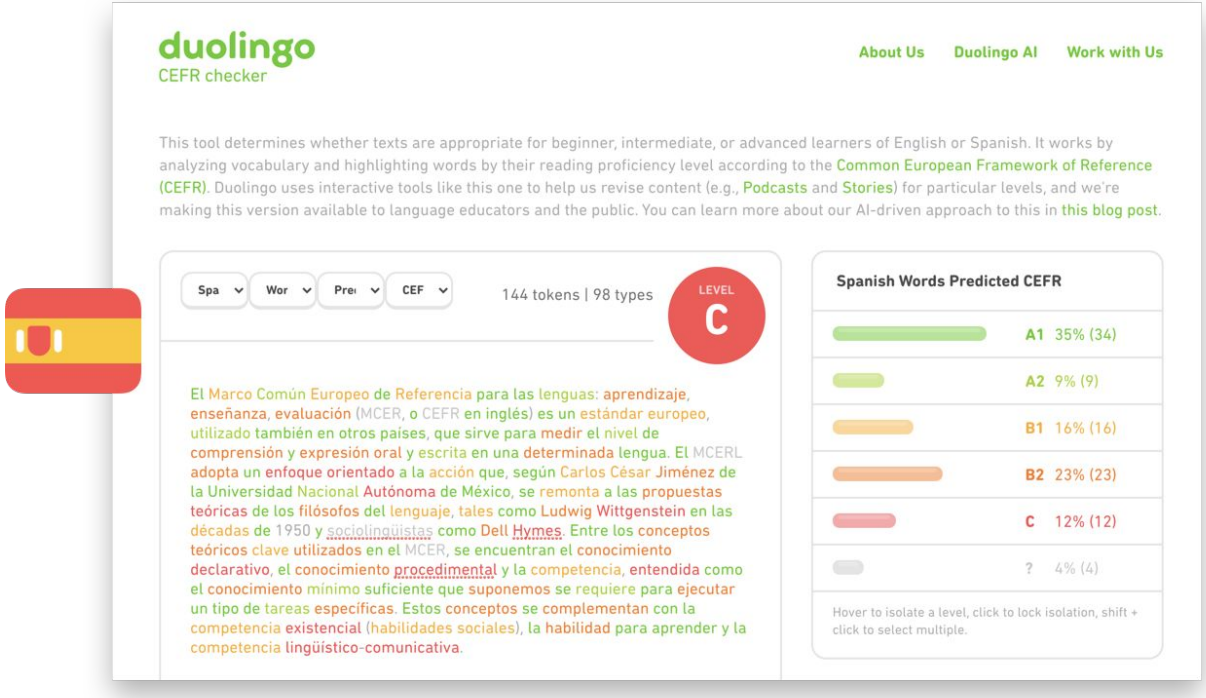
CEFR checker



CEFR checker



CEFR checker



how we've used this at duolingo

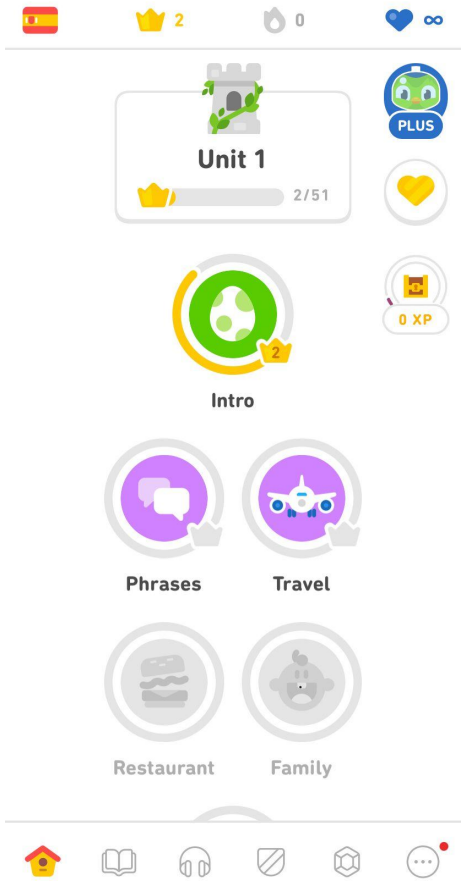
duolingo

- **Mission**— We aim to develop the best education in the world and make it universally available

duolingo

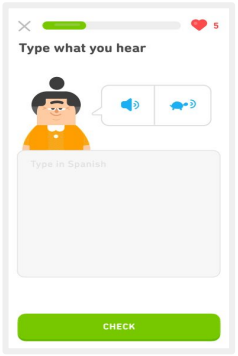
- **Mission**— We aim to develop the best education in the world and make it universally available
- **Current language learning app**
  - Started in 2012
  - More than 500 million users globally
  - Currently language learning 90+ courses (including Irish and Esperanto)
  - Expanding to 100+ courses (including Māori and Yiddish)
  - All learning content is FREE

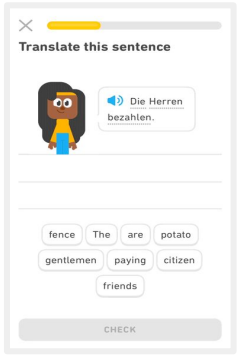
duolingo

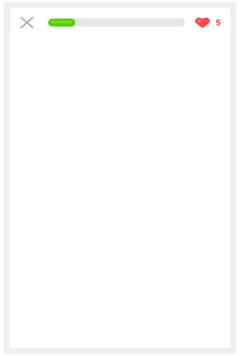

2
0
∞
Unit 1
2/51
PLUS
0 XP
2
Intro
Phrases
Travel
Restaurant
Family

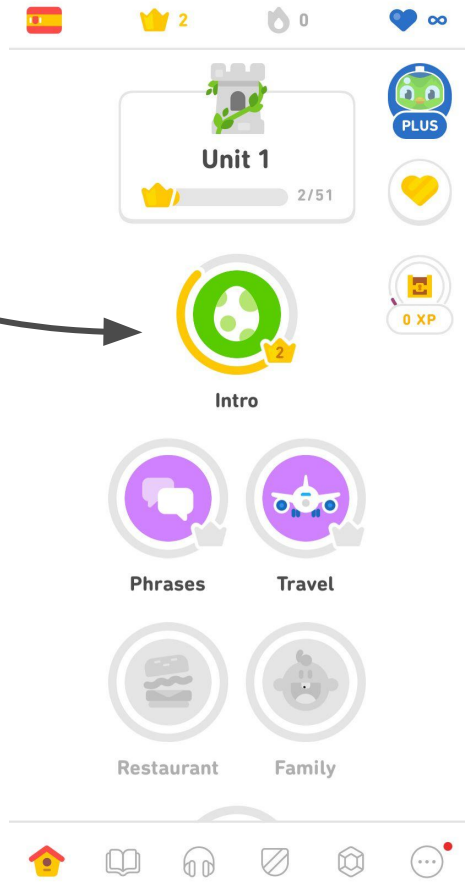
duolingo

game-like lessons grouped into skills by topic



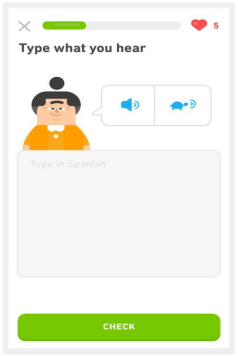


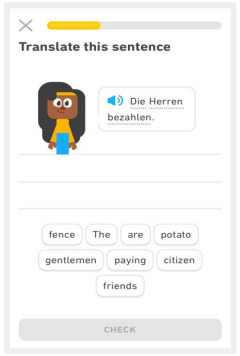


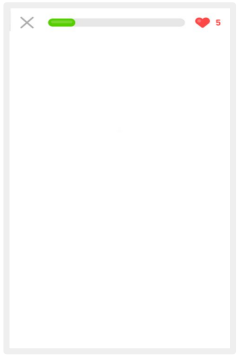


duolingo

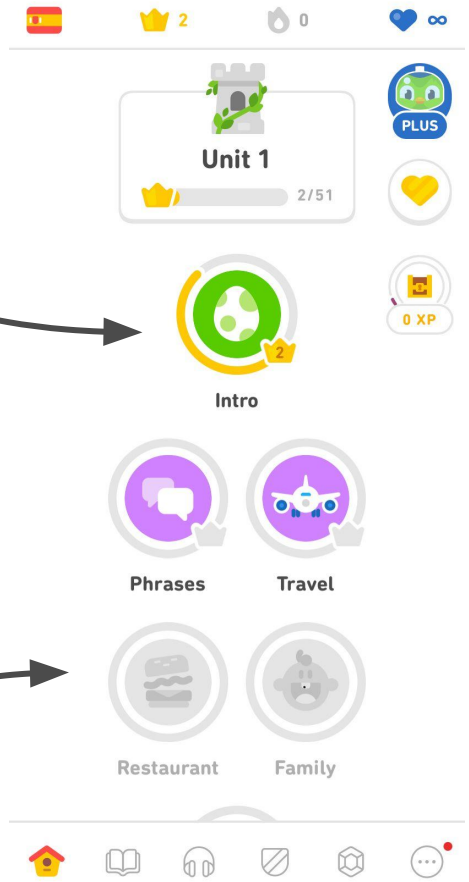
game-like lessons grouped into skills by topic



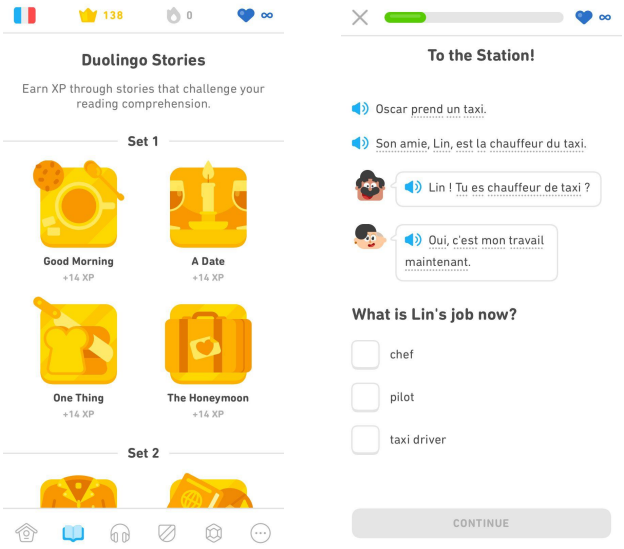




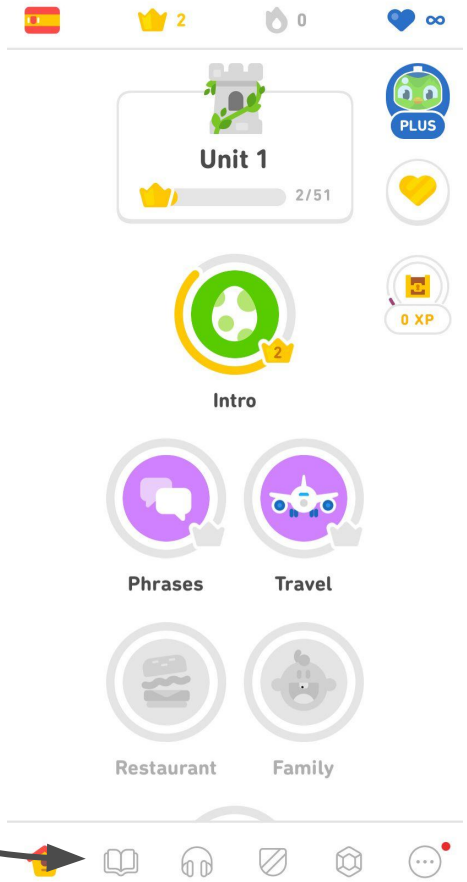
learners complete skills to unlock more rows



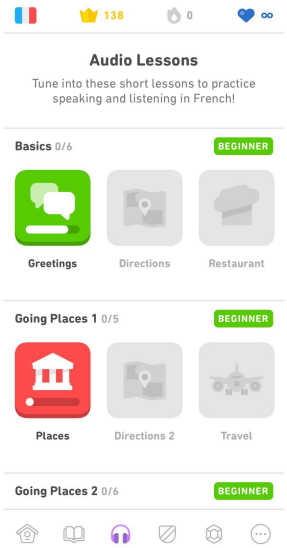
duolingo

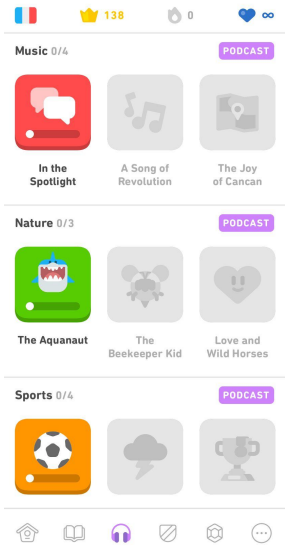


learners can also read interactive stories

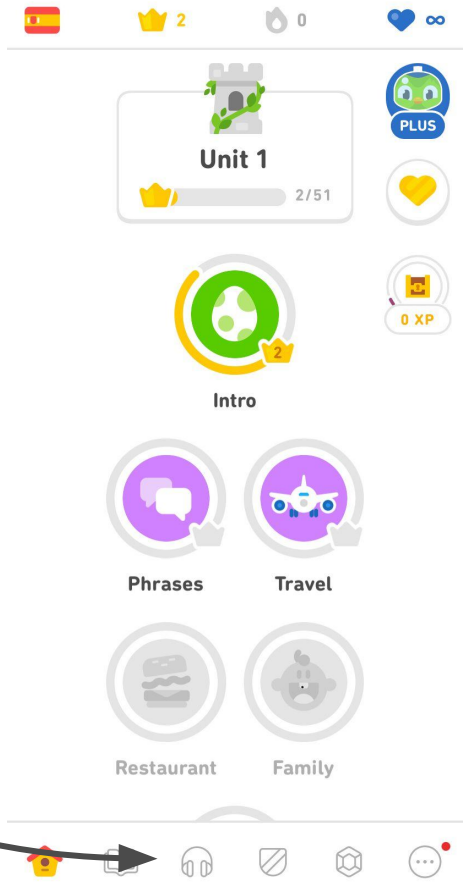


duolingo



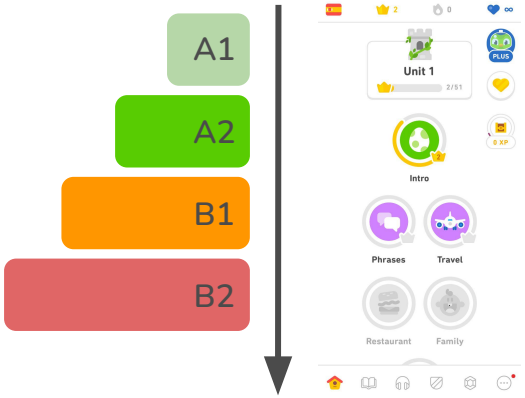


there are also audio lessons and podcasts



CEFR at duolingo

- **Course content** — Courses are aligned the CEFR, such that skills are taught in order by targeted proficiency level

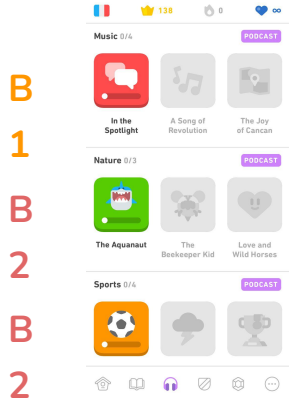


Course content

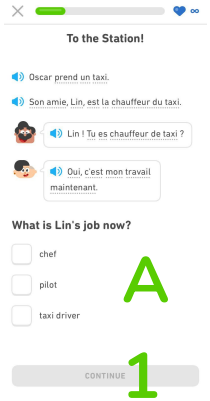
CEFR at duolingo

- **Course content** — Courses are aligned the CEFR, such that skills are taught in order by targeted proficiency level
- **Podcasts and Stories** — Podcast and story content has been adapted to target particular CEFR levels using the CEFR checker

Course content



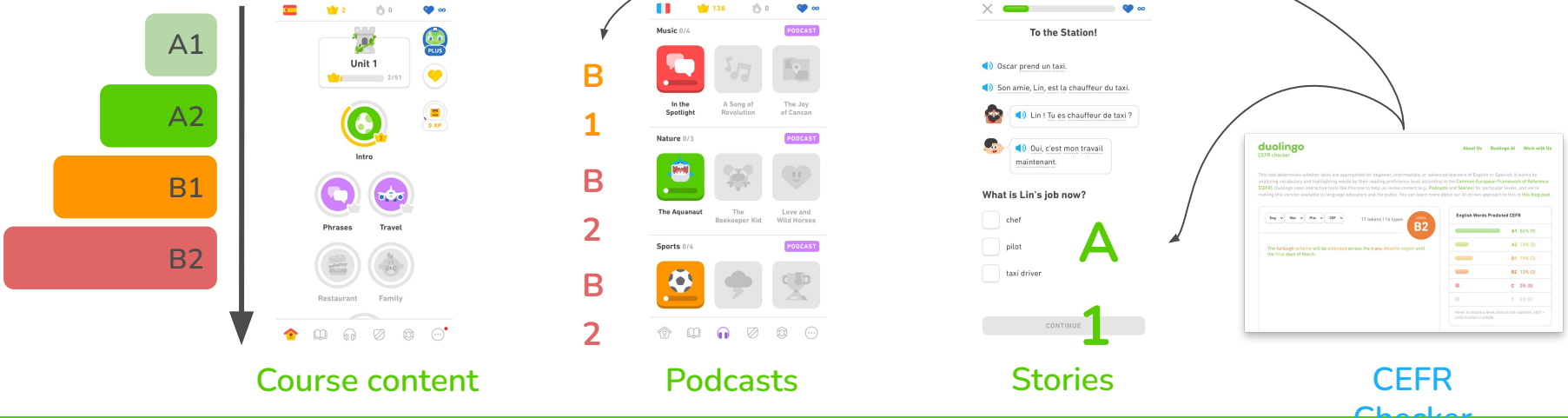
Podcasts



Stories

CEFR at duolingo

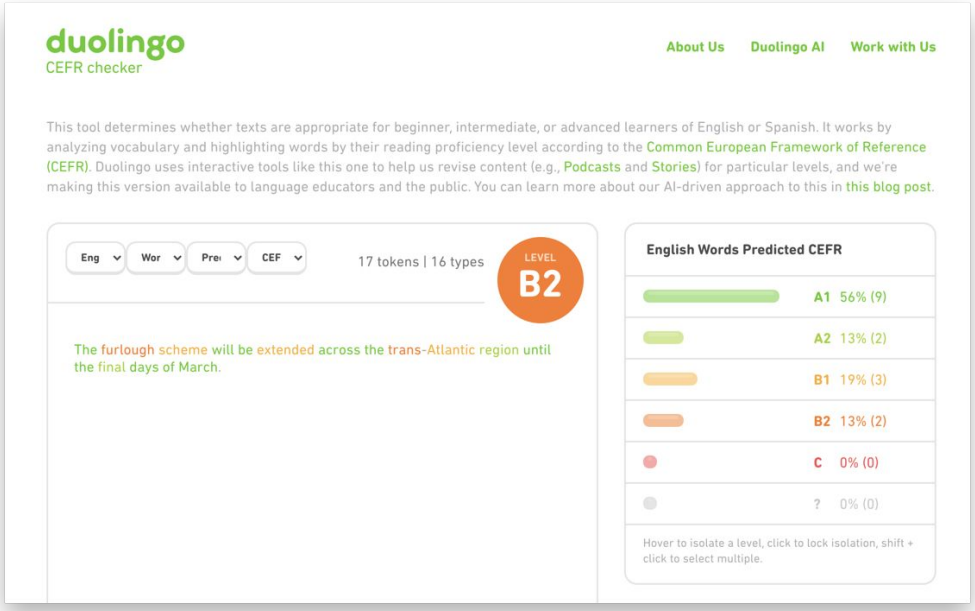
- **Course content** — Courses are aligned the CEFR, such that skills are taught in order by targeted proficiency level
- **Podcasts and Stories** — Podcast and story content has been adapted to target particular CEFR levels using the **CEFR checker**



CEFR checker problem setup

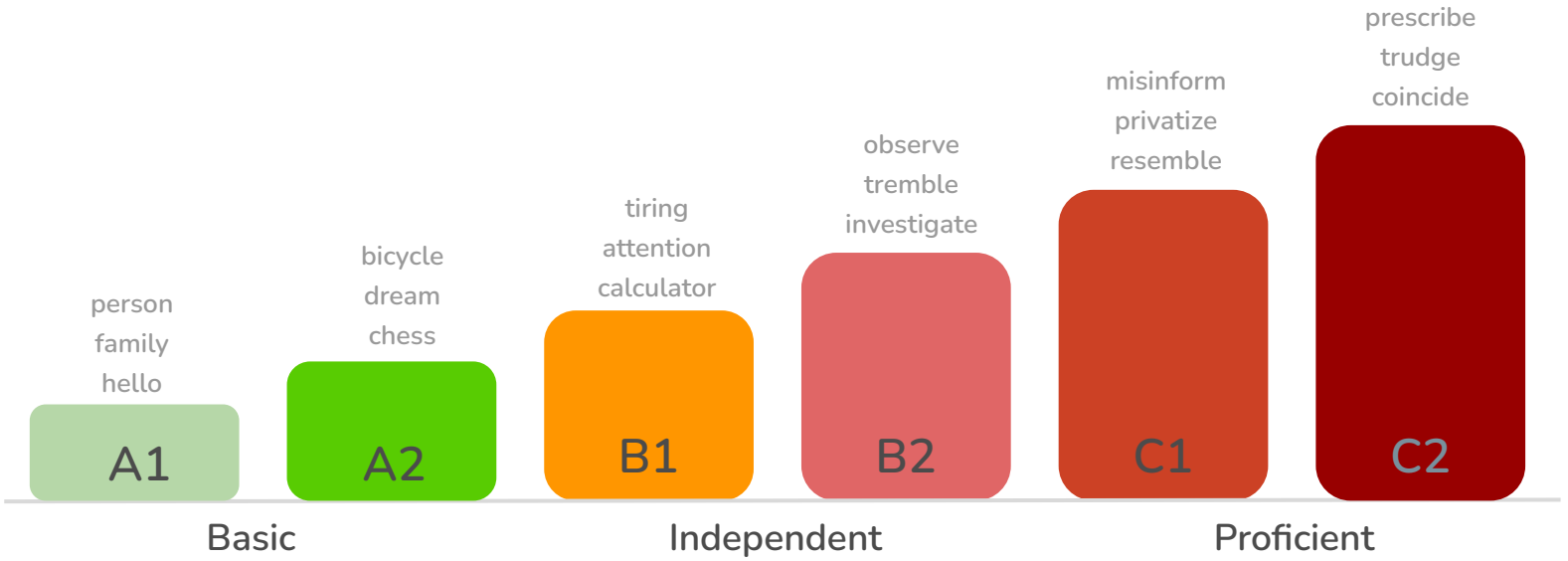
CEFR checker problem setup

- **Want a mapping** — from text to required CEFR proficiency level



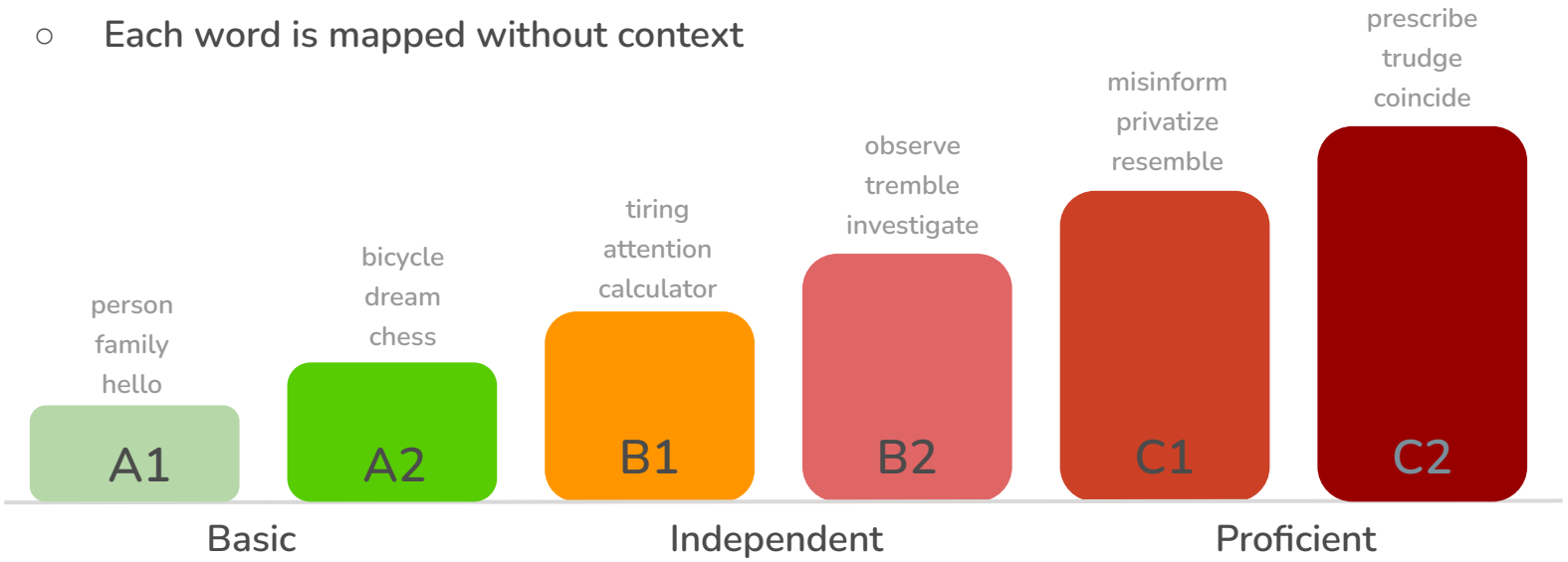
CEFR checker problem setup

- **Want a mapping** — from text to required CEFR proficiency level
- **Simplifications**
  - Map words to CEFR levels



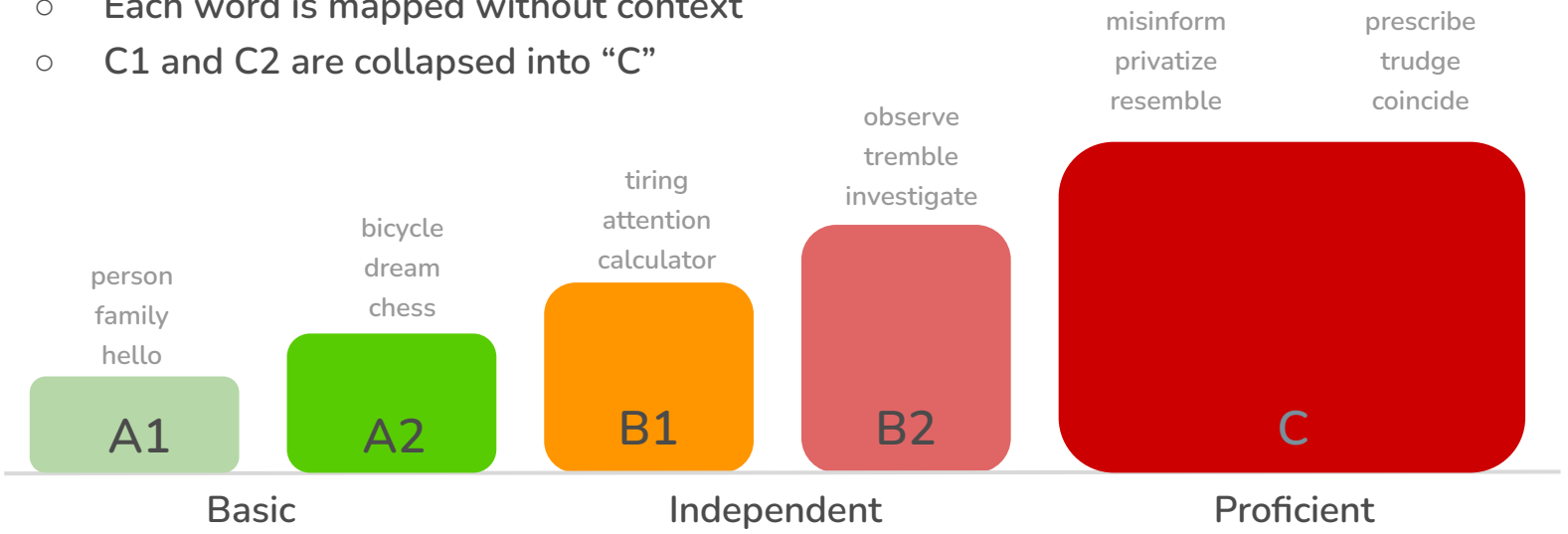
CEFR checker problem setup

- **Want a mapping** — from text to required CEFR proficiency level
- **Simplifications**
  - Map words to CEFR levels
  - Each word is mapped without context



CEFR checker problem setup

- **Want a mapping** — from text to required CEFR proficiency level
- **Simplifications**
  - Map words to CEFR levels
  - Each word is mapped without context
  - C1 and C2 are collapsed into “C”



basic solution

CEFR for words

- **Initial data** — **8800 English words** hand-labeled with CEFR levels
  - Initially labeled based on frequencies across essays from learners at various levels
  - Also further curated internally by Duolingo's curriculum experts
  - Further extended to **5218 Spanish** and **5645 French** words labeled **from A1 up to B1**

CEFR for words

- **Initial data** — **8800 English words** hand-labeled with CEFR levels
  - Initially labeled based on frequencies across essays from learners at various levels
  - Also further curated internally by Duolingo's curriculum experts
  - Further extended to **5218 Spanish** and **5645 French** words labeled **from A1 up to B1**

- **Generalized to other text**— The CEFR tool generalizes from this initial labeled data
  - A wider vocabulary of hundreds of thousands of English, Spanish, and French words
  - Other languages, including German and Italian
    - But these were not available in the public tool

CEFR for words

- **How to generalize?** — Learn a model from the hand-labeled data that maps **language agnostic** word representations to CEFR labels

CEFR for words

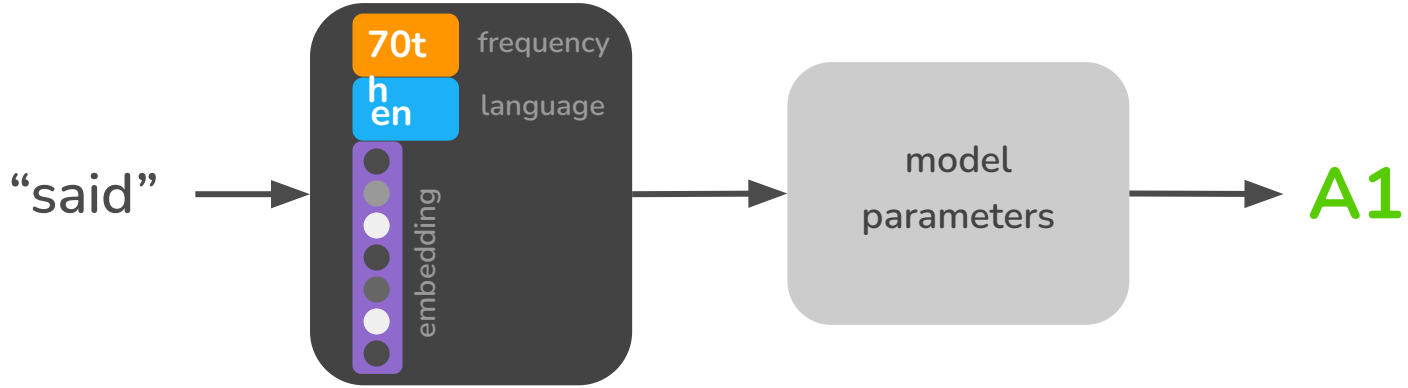
- **How to generalize?** — Learn a model from the hand-labeled data that maps **language agnostic** word representations to CEFR labels
- **Model** — Logistic regression with some minor tweaks

CEFR for words

- **How to generalize?** — Learn a model from the hand-labeled data that maps **language agnostic** word representations to CEFR labels
- **Model** — Logistic regression with some minor tweaks
- **Word representations** — Corpus frequency estimates and multilingual word embeddings (MWEs)

CEFR for words

- **How to generalize?** — Learn a model from the hand-labeled data that maps **language agnostic** word representations to CEFR labels
- **Model** — Logistic regression with some minor tweaks
- **Word representations** — Corpus frequency estimates and multilingual word embeddings (MWEs)



model

- **Multinomial Logistic Regression** — Treat CEFR labels as unordered labels

model

- **Multinomial Logistic Regression** — Treat CEFR labels as unordered labels

- **Ordinal Logistic Regression** — Rennie, Jason DM, and Nathan Srebro. "Loss functions for preference levels: Regression with discrete ordered labels." Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling. Vol. 1. Kluwer Norwell, MA, 2005.
    - Treat CEFR labels as **ordered** labels

model

- **Multinomial Logistic Regression** — Treat CEFR labels as unordered labels

- **Ordinal Logistic Regression** — Rennie, Jason DM, and Nathan Srebro. "Loss functions for preference levels: Regression with discrete ordered labels." Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling. Vol. 1. Kluwer Norwell, MA, 2005.
  - Treat CEFR labels as **ordered** labels
  - Gives a marginal improvement over multinomial logistic regression
    - We generally use this for the results presented here

corpus word frequency estimates

- **OpenSubtitles** — Tiedemann, Jörg. "Parallel data, tools and interfaces in OPUS." Lrec. Vol. 2012. 2012.
  - Word frequencies computed across a large corpus of movie subtitles
  - Over 60 languages and millions of documents

corpus word frequency estimates

- **OpenSubtitles** — Tiedemann, Jörg. "Parallel data, tools and interfaces in OPUS." Lrec. Vol. 2012. 2012.
  - Word frequencies computed across a large corpus of movie subtitles
  - Over 60 languages and millions of documents

- **Features** — We compute several features to allow non-linear relationships to CEFR labels
  - Log raw frequencies
  - Log frequency ranks
  - Bucketed log frequency ranks
    - For several bucketed frequency ranges:
      - 1 if log rank below the range
      - 0 if log rank above the range
      - Between 0 and 1 if log rank in the range

multilingual word embeddings

- **Distributional hypothesis** — Words that share similar contexts share similar meanings

multilingual word embeddings

- **Distributional hypothesis** — Words that share similar contexts share similar meanings

- **Word vectors** — Use the distributional hypothesis to embed words in a vector space such that words with similar meanings (i.e. contexts) are close in the space

multilingual word embeddings

- **Distributional hypothesis** — Words that share similar contexts share similar meanings
- **Word vectors** — Use the distributional hypothesis to embed words in a vector space such that words with similar meanings (i.e. contexts) are close in the space
- **Multilingual embeddings** — Allow words across languages to share the same space, such that words with similar meanings (e.g. translations) are close in the space

multilingual word embeddings

- **MUSE embeddings** — Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)

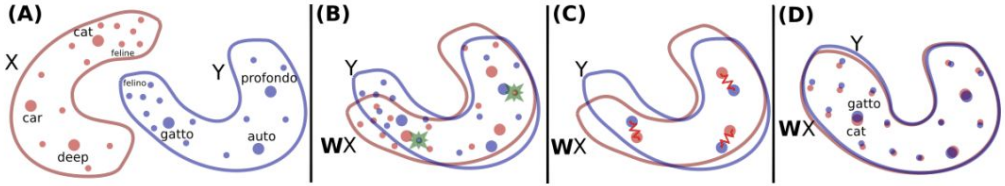


Figure 1: Toy illustration of the method. **(A)** There are two distributions of word embeddings, English words in red denoted by X and Italian words in blue denoted by Y, which we want to align/translate. Each dot represents a word in that space. The size of the dot is proportional to the frequency of the words in the training corpus of that language. **(B)** Using adversarial learning, we learn a rotation matrix W which roughly aligns the two distributions. The green stars are randomly selected words that are fed to the discriminator to determine whether the two word embeddings come from the same distribution. **(C)** The mapping W is further refined via Procrustes. This method uses frequent words aligned by the previous step as anchor points, and minimizes an energy function that corresponds to a spring system between anchor points. The refined mapping is then used to map all words in the dictionary. **(D)** Finally, we translate by using the mapping W and a distance metric, dubbed CSLS, that expands the space where there is high density of points (like the area around the word “cat”), so that “hubs” (like the word “cat”) become less close to other word vectors than they would otherwise (compare to the same region in panel (A)).

experiment: generalizing within languages

training and evaluation

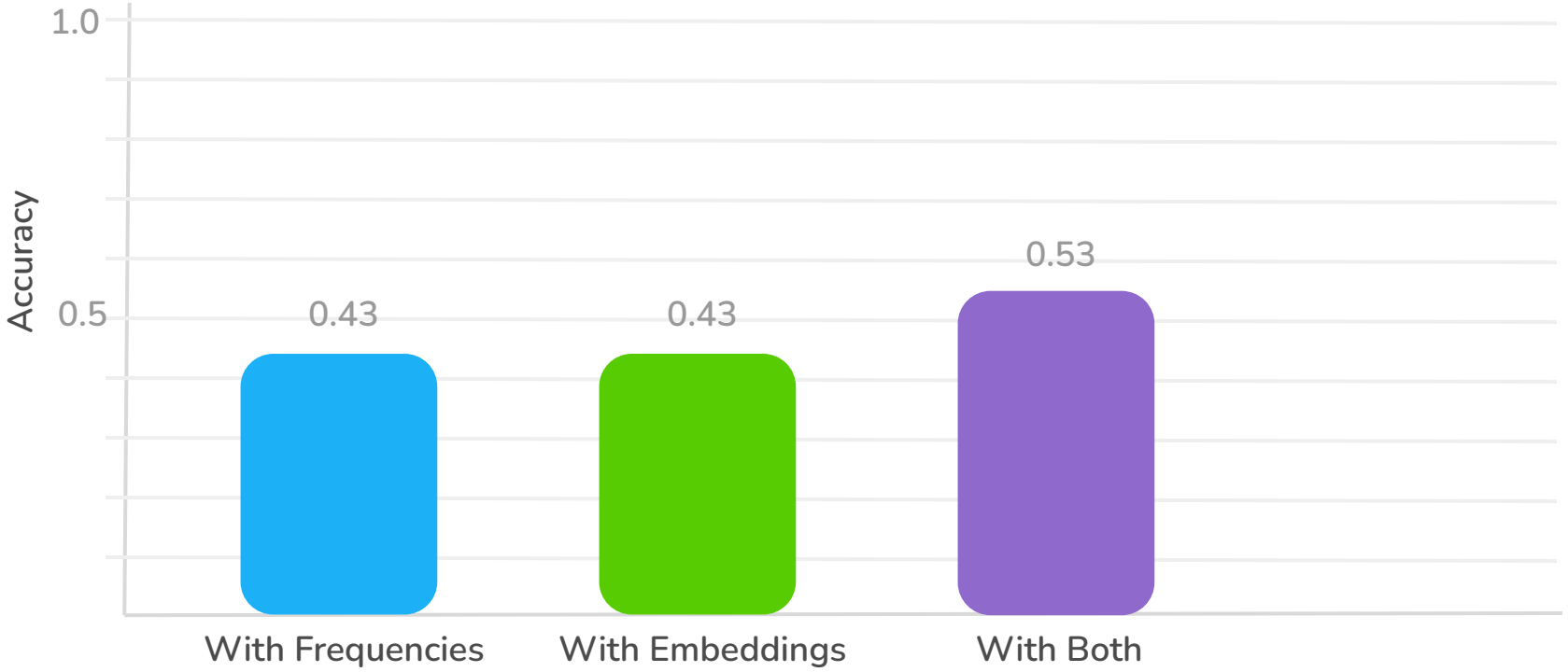
- **Data** — 7226 English, 3764 Spanish, and 3903 French words hand-labeled with CEFR levels
    - Subset of the full labeled data that had MUSE embeddings and OpenSubtitle frequencies
    - French and Spanish only up through B1

training and evaluation

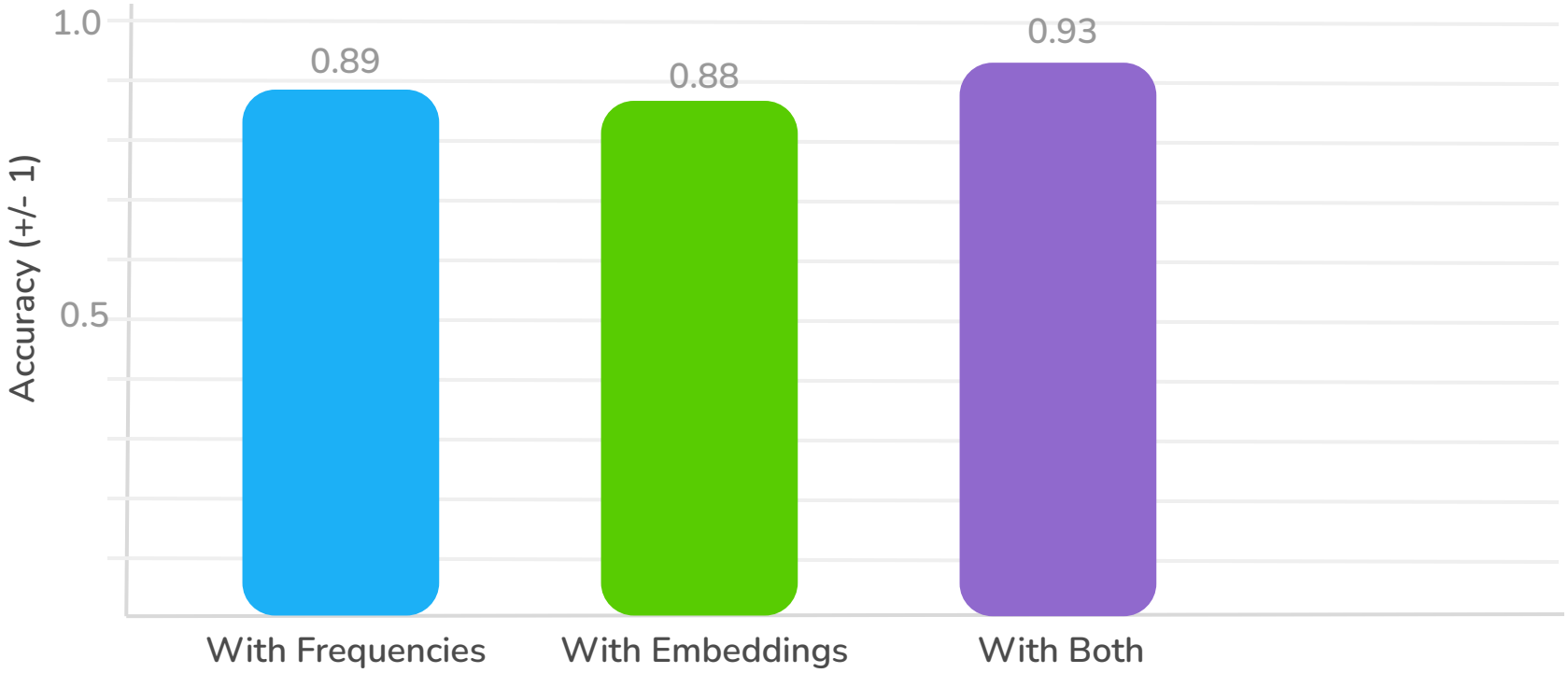
- **Data** — 7226 English, 3764 Spanish, and 3903 French words hand-labeled with CEFR levels
  - Subset of the full labeled data that had MUSE embeddings and OpenSubtitle frequencies
  - French and Spanish only up through B1

- **Evaluation** — 4-fold cross-validation on English, French, and Spanish
  - Can we generalize within these languages?
  - Many evaluations (accuracy, F1, Pearson correlation, Spearman rank correlation, etc)
  - We give accuracies here for simplicity, but all tended to hang together, in general

within language generalization



within language generalization



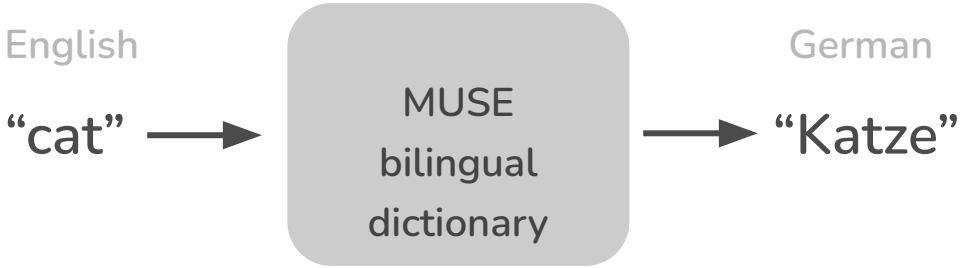
experiment: generalizing to new languages

translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words

translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words
- **Translated data** — Labeled data translated into new languages using bilingual dictionaries
  - Bilingual dictionaries automatically constructed from MUSE embeddings
    - Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)



translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words
- **Translated data** — Labeled data translated into new languages using bilingual dictionaries
  - Bilingual dictionaries automatically constructed from MUSE embeddings
    - Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)

translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words
- **Translated data** — Labeled data translated into new languages using bilingual dictionaries
  - Bilingual dictionaries automatically constructed from MUSE embeddings
    - Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)
  - Translating from English CEFR labeled words, we construct:
    - 8469 Italian CEFR labeled words
    - 8755 German CEFR labeled words
    - 4153 Spanish B2 and C level CEFR labeled words
    - 4161 French B2 and C level CEFR labeled words

translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words
- **Translated data** — Labeled data translated into new languages using bilingual dictionaries
  - Bilingual dictionaries automatically constructed from MUSE embeddings
    - Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)
  - Translating from English CEFR labeled words, we construct:
    - 8469 Italian CEFR labeled words
    - 8755 German CEFR labeled words
    - 4153 Spanish B2 and C level CEFR labeled words
    - 4161 French B2 and C level CEFR labeled words
- **Can we generalize from labeled to translated data?** — Evidence for several hypotheses:
  - Does translating using bilingual dictionaries consistently transfer CEFR labels?
  - Can our models generalize across languages?

translating data across languages

- **Labeled data** — CEFR labeled English, French (up to B1), and Spanish (up to B1) words
- **Translated data** — Labeled data translated into new languages using bilingual dictionaries
  - Bilingual dictionaries automatically constructed from MUSE embeddings
    - Conneau, Alexis, et al. “Word translation without parallel data.” arXiv preprint arXiv:1710.04087 (2017)
  - Translating from English CEFR labeled words, we construct:
    - 8469 Italian CEFR labeled words
    - 8755 German CEFR labeled words
    - 4153 Spanish B2 and C level CEFR labeled words
    - 4161 French B2 and C level CEFR labeled words
- **Can we generalize from labeled to translated data?** — Evidence for several hypotheses:
  - Does translating using bilingual dictionaries consistently transfer CEFR labels?
  - Can our models generalize across languages?

note missing simpler evaluation

cross language generalization

- **Generalize across languages?** — Can a model trained on original labeled data generalize to translated data?

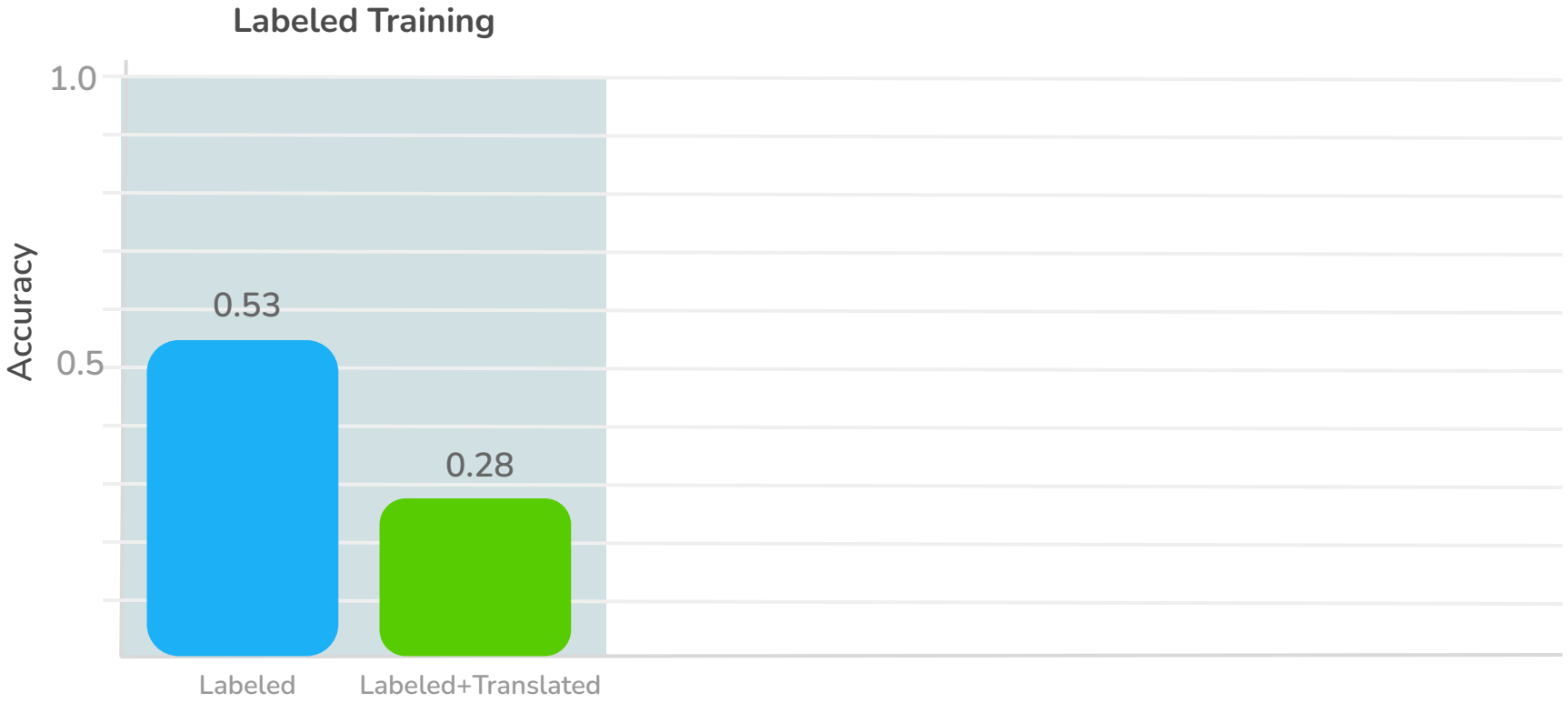
cross language generalization

- **Generalize across languages?** — Can a model trained on original labeled data generalize to translated data?
    - Features are language agnostic (corpus frequencies and MWEs)

cross language generalization

- **Generalize across languages?** — Can a model trained on original labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments

cross language generalization



cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:

cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:
    i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages

cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:
    i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages
    ii. Model features are not language agnostic from the perspective of a linear model

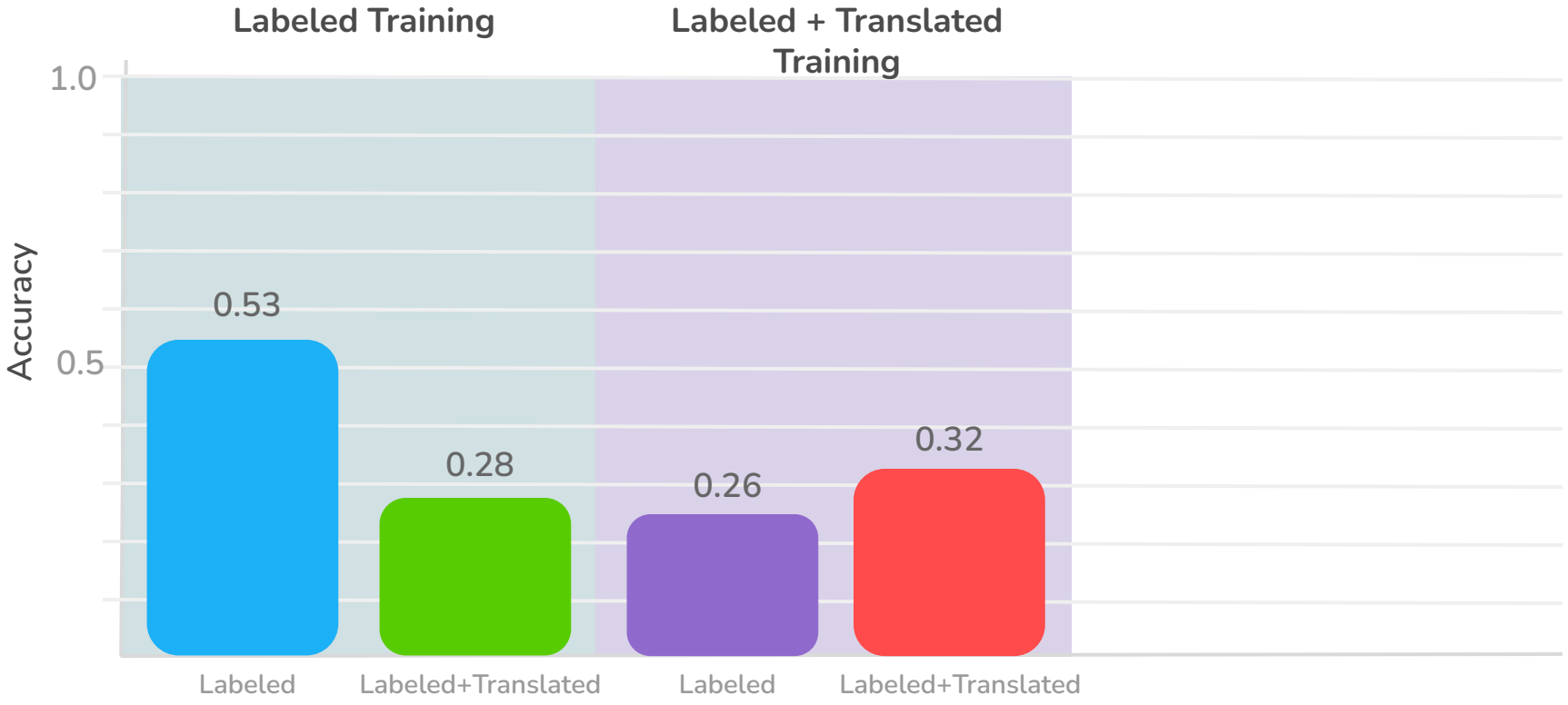
cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:
    i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. Translating labels across languages does not produce consistent labelings

cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:
    i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. Translating labels across languages does not produce consistent labelings
- **Generalize within translated data** — Can a linear model trained on all labeled and translated data generalize to unseen labeled and translated data?

cross language generalization



cross language generalization

- **Generalize across languages** — Can a model trained on labeled data generalize to translated data?
  - Features are language agnostic (corpus frequencies and MWEs)
  - Evaluate with 4-fold cross validation over translated data, similar to within language generalization experiments
  - Accuracy on translated data is lower, so something is wrong. Possibilities:
    i. ~~Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. Translating labels across languages does not produce consistent labelings
- **Generalize within translated data** — Can a linear model trained on all labeled and translated data generalize to unseen labeled and translated data?

cross language adaptation

- **Domain adaptation** — Daumé III, Hal. "Frustratingly Easy Domain Adaptation." ACL 2007 (2007): 256.

cross language adaptation

- **Domain adaptation** — Daumé III, Hal. "Frustratingly Easy Domain Adaptation." ACL 2007 (2007): 256.
    - Keep all existing language-agnostic corpus frequency and MWE features, but add an additional copy of them for each language

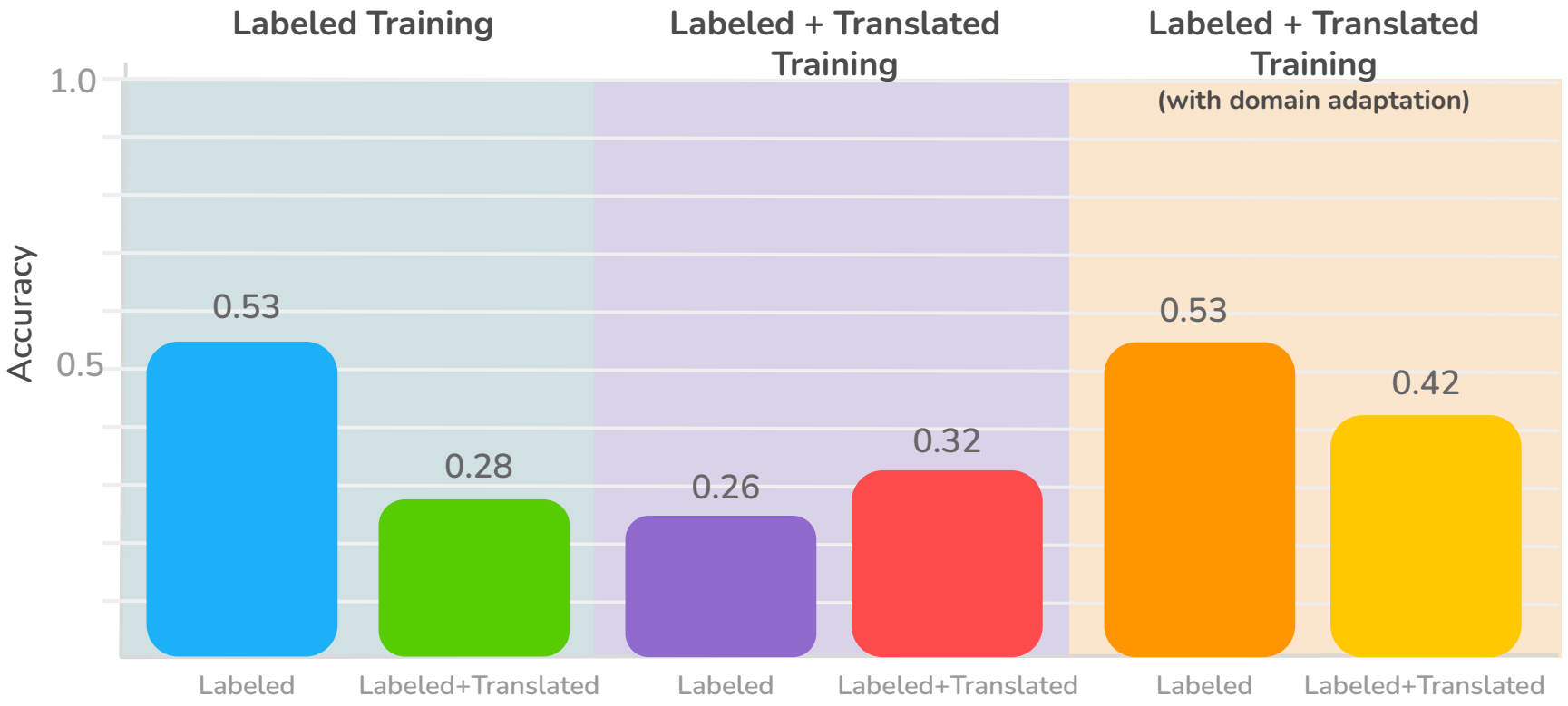
cross language adaptation

- **Domain adaptation** — Daumé III, Hal. "Frustratingly Easy Domain Adaptation." ACL 2007 (2007): 256.
  - Keep all existing language-agnostic corpus frequency and MWE features, but add an additional copy of them for each language
    - Language-specific features copy original feature values for words of particular language, but are zero valued for all other languages

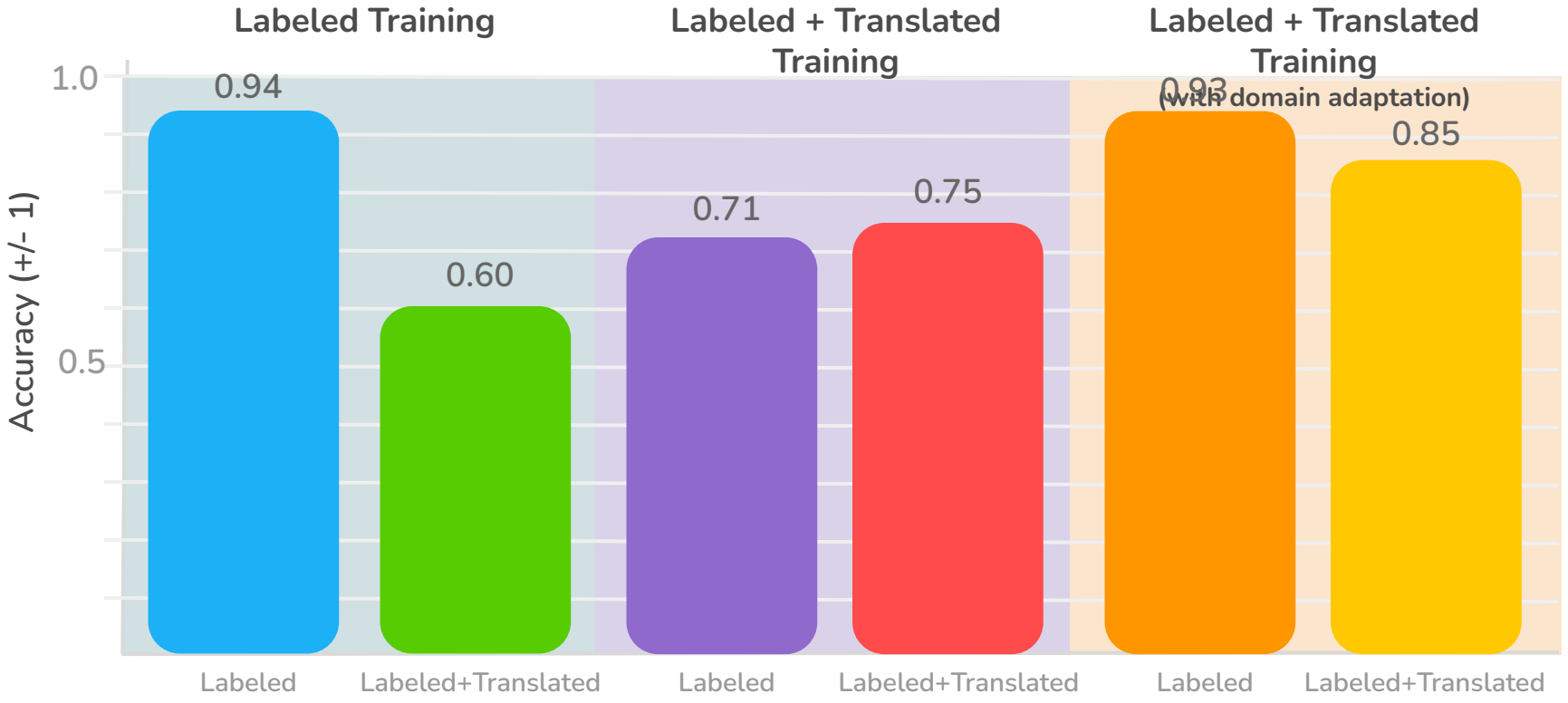
cross language adaptation

- **Domain adaptation** — Daumé III, Hal. "Frustratingly Easy Domain Adaptation." ACL 2007 (2007): 256.
  - Keep all existing language-agnostic corpus frequency and MWE features, but add an additional copy of them for each language
    - Language-specific features copy original feature values for words of particular language, but are zero valued for all other languages
- **Generalize with domain adaptation?** — Then translating labels across languages produces consistent labelings, but features are not language agnostic with respect to a linear model

cross language adaptation



cross language adaptation (+/- 1)



cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
    - For our original set of hypotheses about why a non-adapted model doesn't generalize:

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
    - For our original set of hypotheses about why a non-adapted model doesn't generalize:
        i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages
        ii. Model features are not language agnostic from the perspective of a linear model
        iii. Translating labels across languages does not produce consistent labelings

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
  - For our original set of hypotheses about why a non-adapted model doesn't generalize:
    i. ~~Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. ~~Translating labels across languages does not produce consistent labelings~~

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
    - For our original set of hypotheses about why a non-adapted model doesn't generalize:
        i. ~~Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
        ii. Model features are not language agnostic from the perspective of a linear model
        iii. ~~Translating labels across languages does not produce consistent labelings~~
- **For our original questions** — Can we generalize to translated data?

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
  - For our original set of hypotheses about why a non-adapted model doesn't generalize:
    i. ~~Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. ~~Translating labels across languages does not produce consistent labelings~~
- **For our original questions** — Can we generalize to translated data?
  - Does translating using bilingual dictionaries consistently transfer CEFR labels? Probably

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
  - For our original set of hypotheses about why a non-adapted model doesn't generalize:
    i. ~~Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
    ii. Model features are not language agnostic from the perspective of a linear model
    iii. ~~Translating labels across languages does not produce consistent labelings~~
- **For our original questions** — Can we generalize to translated data?
  - Does translating using bilingual dictionaries consistently transfer CEFR labels? Probably
  - Can our models generalize across languages? Not amazingly well, but a little

cross language adaptation

- **We can generalize with domain adaptation, but not without it** — Labels are consistent, but features are not very linearly language agnostic with respect to CEFR labels
  - For our original set of hypotheses about why a non-adapted model doesn't generalize:
    - ~~i. Features are language agnostic from the perspective of a linear model, but model parameters are not well-tuned to generalize across languages~~
    - ii. Model features are not language agnostic from the perspective of a linear model
    - ~~iii. Translating labels across languages does not produce consistent labelings~~
- **For our original questions** — Can we generalize to translated data?
  - Does translating using bilingual dictionaries consistently transfer CEFR labels? Probably
  - Can our models generalize across languages? Not amazingly well, but a little

other things we tried

other experiments

- **Joint pairwise and ordinal model** — Sculley, David. "Combined regression and ranking." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010.
    - Can we learn if we have many pairwise labels, but only a few ordinal labels?
    - Under domain adaptation, original labeled ordinals but translated data only as pairs performs almost as well as model with translated ordinals

other experiments

- **Joint pairwise and ordinal model** — Sculley, David. "Combined regression and ranking." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010.
    - Can we learn if we have many pairwise labels, but only a few ordinal labels?
    - Under domain adaptation, original labeled ordinals but translated data only as pairs performs almost as well as model with translated ordinals
- CEFRLex based labels — https://cental.uclouvain.be/cefrlex/bibliography/
    - For simplicity, only used our internal labels eventually, but there may have been some other ways we could have used these

other experiments

- **Joint pairwise and ordinal model** — Sculley, David. "Combined regression and ranking." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010.
    - Can we learn if we have many pairwise labels, but only a few ordinal labels?
    - Under domain adaptation, original labeled ordinals but translated data only as pairs performs almost as well as model with translated ordinals
- CEFRLex based labels — https://cental.uclouvain.be/cefrlex/bibliography/
    - For simplicity, only used our internal labels eventually, but there may have been some other ways we could have used these
- Googlebooks — https://books.google.com/ngrams
    - We used Googlebook word frequency estimates in addition to OpenSubtitles
    - OpenSubtitles tended to produce better results

possible future work

future work

- **Contextual predictions** — Use large language model (e.g. BERT) embeddings to predict CEFR labels for words in context

future work

- **Contextual predictions** — Use large language model (e.g. BERT) embeddings to predict CEFR labels for words in context
- **More languages** — Extend these methods to produce predictions across additional languages

questions?

thank you