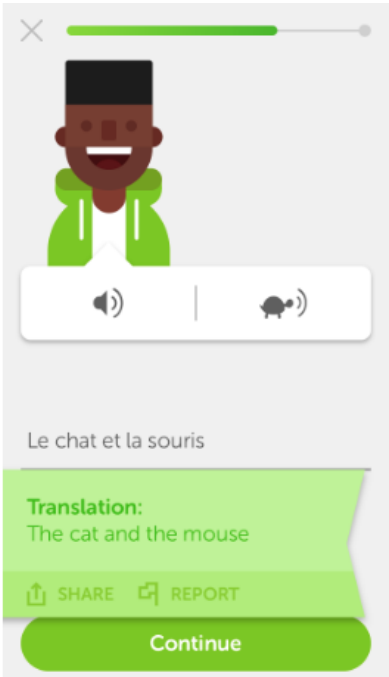
2018 Duolingo SLAM（Second Language Acquisition Modeling ）

背景：

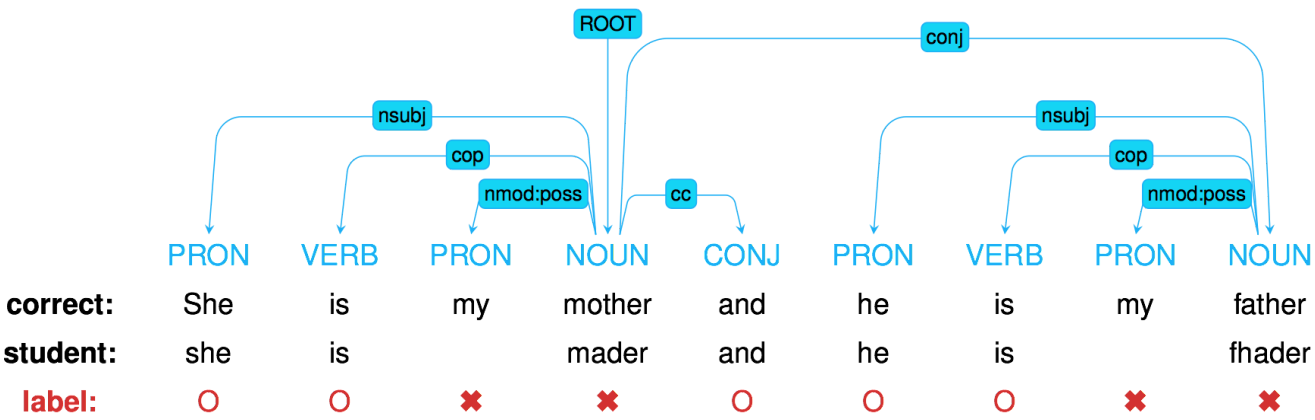
Duolingo在线语言学习平台下的三种互动式的练习形式

- reverse_translate：学生阅读他们所熟悉的语言材料，并翻译成他们正在学习的语言
- reverse_tap：学生从一堆词汇中选出翻译结果
- listen：听一段话（正在学习的语言），并进行转录



目标：

预测英语、西班牙语和法语的学习者将来会犯的错误，这些错误都是基于他们过去犯过的错误



训练集中 0表示完美 1表示错误（忽略大写、标点符号、口音）

数据集内容：（可下载）

包含了超过200万个来自超过6000名 Duolingo 学生在他们的前30天提交的答案，允许使用外部数据

```
# user:D2inSf5+  countries:MX  days:1.793  client:web  session:lesson  format:reverse_translate  time:16
8rgJEAPw1001  She          PRON    Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP    nsubj       4  0
8rgJEAPw1002  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         4  0
8rgJEAPw1003  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   4  1
8rgJEAPw1004  mother       NOUN    Degree=Pos|fPOS=ADJ++JJ                                                ROOT        0  1
8rgJEAPw1005  and          CONJ    fPOS=CONJ++CC                                                          cc          4  0
8rgJEAPw1006  he           PRON    Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs|fPOS=PRON++PRP   nsubj       9  0
8rgJEAPw1007  is           VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBZ    cop         9  0
8rgJEAPw1008  my           PRON    Number=Sing|Person=1|Poss=Yes|PronType=Prs|fPOS=PRON++PRP$              nmod:poss   9  1
8rgJEAPw1009  father       NOUN    Number=Sing|fPOS=NOUN++NN                                              conj        4  1

# user:D2inSf5+  countries:MX  days:2.689  client:web  session:practice  format:reverse_translate  time:6
oMGsnnH/0101  When         ADV     PronType=Int|fPOS=ADV++WRB                                             advmod      4  1
oMGsnnH/0102  can          AUX     VerbForm=Fin|fPOS=AUX++MD                                              aux         4  0
oMGsnnH/0103  I            PRON    Case=Nom|Number=Sing|Person=1|PronType=Prs|fPOS=PRON++PRP              nsubj       4  1
oMGsnnH/0104  help         VERB    VerbForm=Inf|fPOS=VERB++VB                                             ROOT        0  0
```

三种竞赛轨道：

- `en_es` — **English learners 英语学习者** (已经会说西班牙语)
- `es_en` — **Spanish learners 西班牙语学习者**(已经会说英语)
- `fr_en` — **French learners 法语学习者** (已经会说英语)

评价方法：AUC(主要指标) F1

结果提交文件：针对每一个练习实例，输出学生犯错的可能性

```
DRihrVmh0901  0.025
DRihrVmh0902  0.08
DRihrVmh0903  0.454
DRihrVmh0904  0.044
TOeLHxLS0401  0.067
TOeLHxLS0402  0.03
TOeLHxLS0403  0.806
TOeLHxLS0404  0.066
xqtN1I5c0901  0
xqtN1I5c0902  0.074
xqtN1I5c0903  0.053
xqtN1I5c0904  0.016
...
```

官方解决方案参考：

- IRT模型（Item Response Theory）
- Modeling Learning & Forgetting Over Time
  - BKT: HMM,估计一个学生什么时候学会了一个概念
  - DKT: 使用RNN替代BKT中的HMM
  - Half-life Regression: 指数型"遗忘曲线"的机器学习模型
  - 100 Years of Forgetting: 学习不同的遗忘曲线
- 语言规则：
  - 语言形态学规则

    - 句法规则
- 语法错误检测：CRF 、HMM、RNN
- 多语言学习： 把三种语言任务合并为一个

冠军解决方案（先声教育）：

提出了一种CLUF网络：

- Context encode： 字层级LSTM + 词层级LSTM
- Linguistic feature encoder（语法解析）： 2层LSTM
- User information encoder（用户属性）： 1层DNN
- Format information encode （练习实例属性）: 1层DNN
- decoder： sigmoid