

Watch Drama, Learn Language

Liling Tan

Rakuten Institute of Technology (Singapore)

10 Jan 2018 @ Global WordNet Conference



Overview

- **Demos**
- **Generate Quiz Automatically**
- **Numbers and Graphs**
- **WordNet and Language Learning**

Demos

Drama Vocab Quiz

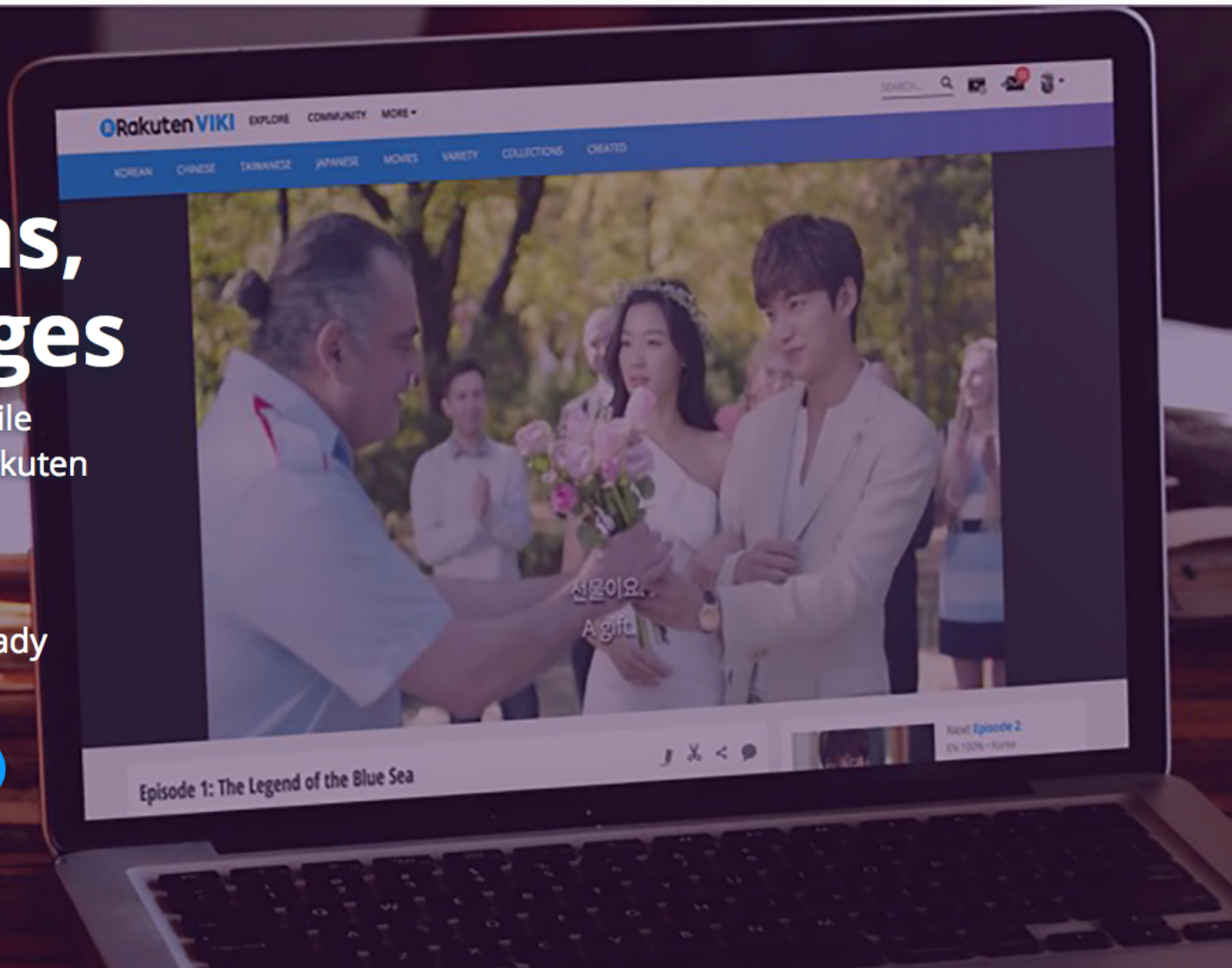
Watch dramas, learn languages

Learn Korean and Chinese words while
watching your favorite dramas on Rakuten
Viki!

How many drama words do you already
know? Take the quiz to find out!


KOREAN QUIZ

CHINESE QUIZ



1 / 30

전화

jeon-hwa 

anticipation


signal

telephone

I don't know >

8 / 30

낭만

nang-man 

logical

romance

soldier

I don't know >

 Report issues

Your results



Hey,

Heol! (Omg!) You know about **11%** of the common drama words. It's a start, but you need to try harder to impress Jisung.



[SEE DETAILED RESULTS](#)

Share your result with friends. See who scores better!



Results Breakdown



30 **7** **17**

Total Correct Wrong

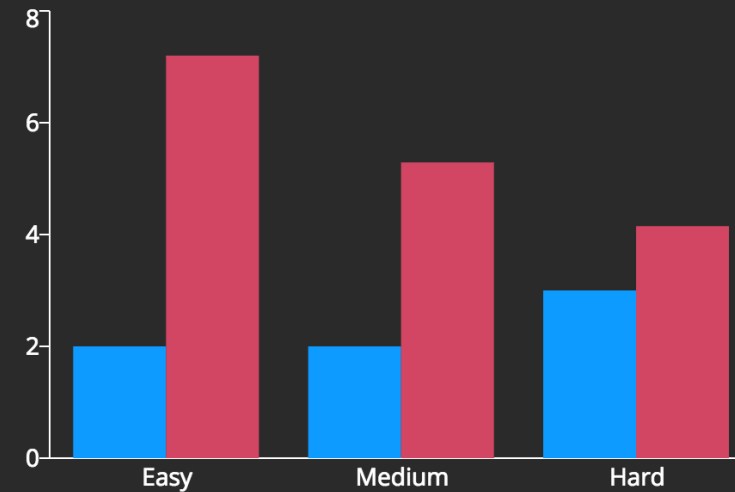
Out of the 30 words in the quiz, there were 10 easy, 10 medium, and 10 hard words. You answered **2 easy** words, **2 medium** words and **3 hard** words **correctly**, and **17 words incorrectly**. Based on your answers, we estimate that you know about **1%** of the common drama words.

Please note that getting answers wrong decreases your score. If you are unsure of the answer, you can **select** the "**I don't know**" option - this will not impact your score negatively.

You think you can do better? Try again! Every time you take our quiz, you see a new set of words.

Curious to know the science behind your score? Find out [here](#).

● Words you got correct ● Words others got correct



Watch and Learn

Here are some of the words you missed. With Learn mode, you can watch dramas and learn Korean at the same time. Check it out!

WATCH AND LEARN



유서 yu-seo History

This word appears in

EP 1: The Legend of the Blue Sea



낭만 nang-man Romance

This word appears in

EP 9: Healer

SEE MORE

Generate Quiz Automatically



Select a list of words
and translation

자격증

Certificate

Document

Church

I don't know

Select distractors

How to generate quiz automatically?

- **Select a list of words**
 - Selected sample needs to be representative of the corpus
- **Find the translation**
 - Translation has to be accurate and most prototypical / least ambiguous
- **Generate distractors**
 - Distractors MUST NOT have the same meaning as the answer
 - Distractors should not be too easy to eliminate

Select a list of words

1. **Create a Corpus:** Collect all captions from Korean dramas
2. **Tokenize:** Extract lemmas from surface words.

미남	- 이(다)	- (으)시 -	- (이)라구	- 요
minam	- is	- si -	- lagu	- yo
handsome guy	(he) is	(honorific)	Someone said	(polite marker)

- The word 미남 [minam] “handsome guy” can be followed by multiple suffixes at once -이시라구요 [-issilaguyo] to form a single word meaning “Someone said that he is handsome”.
- Extract the root word 미남 [minam], and count it as a unique word type

Select a list of words

1. Create a Corpus

2. Tokenize

3. Filter Rare Words

- Term Frequency > 20
- Drama Frequency > 5

Word	English
송전	power transmission
림프절	lymph nodes
감가상각	depreciation
달맞이꽃	evening primrose
보리밭	barley field
열화학	thermal chemistry
식기세척기	dishwasher
테라포밍	terraforming
삼각함수	trigonometric function
지방산	fatty acid

Select a list of words

1. Create a Corpus
2. Tokenize
3. Filter Rare Words

4. Translate

- Find translations from **in-house** dictionary and MT engines.
- **Dictionaries:** More info on slides 12 and 15-38
<https://www.slideshare.net/rakutentech/ai-based-language-learning-tools>
- **Machine Translation:** Come talk to me =)

Select a list of words

1. Create a Corpus
2. Tokenize
3. Filter Rare Words
4. Translate
- 5. Filter More Words**
 - Konglish words are too easy to guess
 - Swear words are inappropriate
 - <https://github.com/alvations/expletives>

Filter Konglish Words

Korean
그린
배드민턴
스트라이크
홍혜정
햄버거
보디가드
김태평
설렁탕
카피라이터
패션모델

English
green
badminton
strike
hong hye jung
hamburger
bodyguard
kim tae pyung
seolleongtang
copywriter
fashion model

Filter Konglish Words

Korean	Romanization
그린	geu-rin
배드민턴	bae-deu-min-teon
스트라이크	seu-teu-ra-i-keu
홍혜정	hong-hye-jeong
햄버거	haem-beo-geo
보디가드	bo-di-ga-deu
김태평	gim-tae-pyeong
설렁탕	seor-reong-tang
카피라이터	ka-pi-ra-i-teo
패션모델	pae-syeon-mo-der

English
green
badminton
strike
hong hye jung
hamburger
bodyguard
kim tae pyung
seolleongtang
copywriter
fashion model

Filter Konglish Words

Korean	Romanization	Romanization (No vowel)
그린	geu-rin	grn
배드민턴	bae-deu-min-teon	bdmntn
스트라이크	seu-teu-ra-i-keu	strk
홍혜정	hong-hye-jeong	hng hy jng
햄버거	haem-beo-geo	hmbrgr
보디가드	bo-di-ga-deu	bdgd
김태평	gim-tae-pyeong	gmtpyng
설렁탕	seor-reong-tang	srrngtng
카피라이터	ka-pi-ra-i-teo	kprrt
패션모델	pae-syeon-mo-der	psynmdr

English (No vowel)	English
grn	green
bdmntn	badminton
strk	strike
hng hy jng	hong hye jung
hmbrgr	hamburger
bdygrd	bodyguard
km t pyng	kim tae pyung
sllngtng	seolleongtang
cpywrtr	copywriter
fshn mdl	fashion model

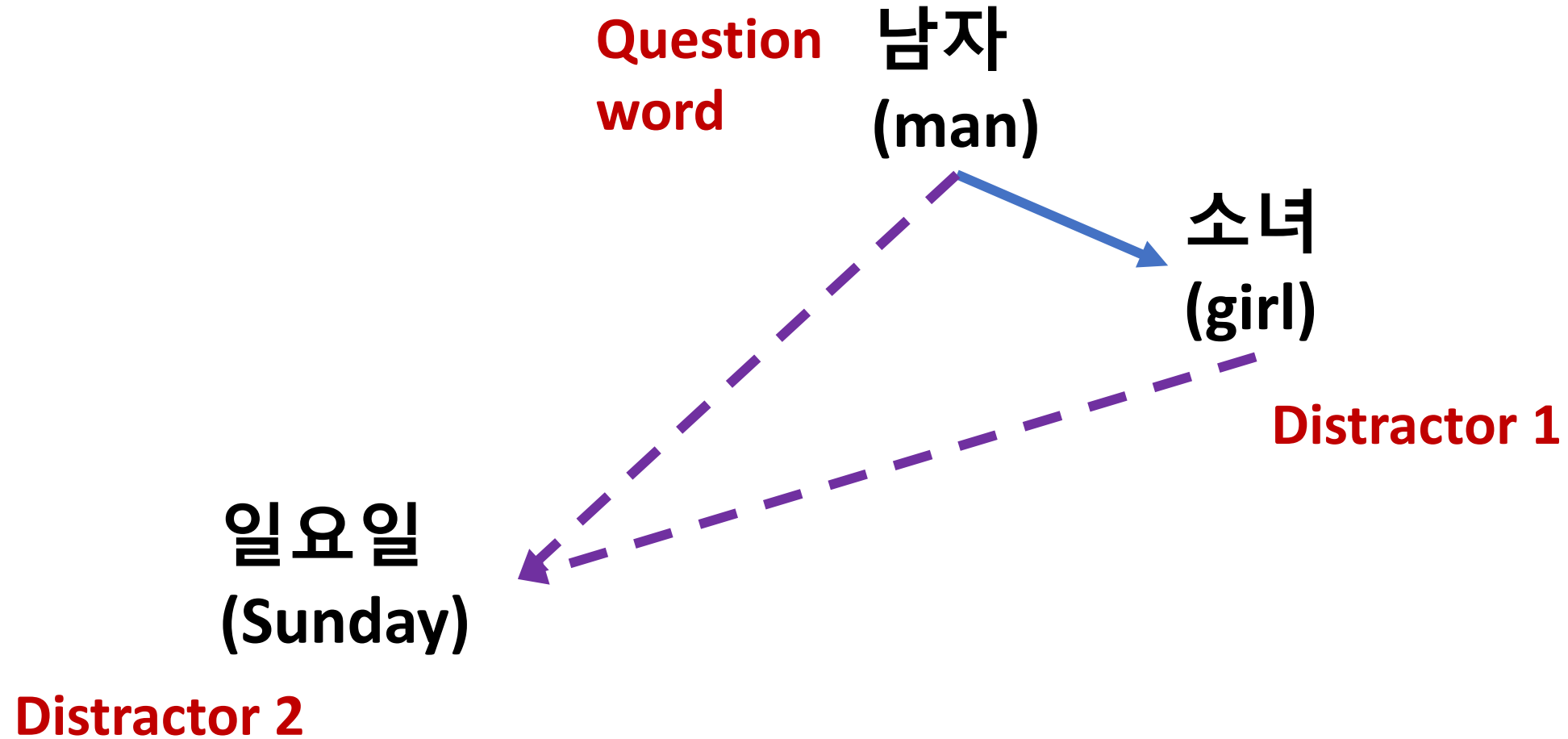
Filter Konglish Words

Korean	Romanization	Romanization (No vowel)	Fuzzy Score (Levenshtein)	English (No vowel)	English
그린	geu-rin	grn	100	grn	green
배드민턴	bae-deu-min-teon	bdmntn	100	bdmntn	badminton
스트라이크	seu-teu-ra-i-keu	strk	100	strk	strike
홍혜정	hong-hye-jeong	hnghyjng	89	hng hy jng	hong hye jung
햄버거	haem-beo-geo	hmbg	80	hmbrgr	hamburger
보디가드	bo-di-ga-deu	bdgd	80	bdygrd	bodyguard
김태평	gim-tae-pyeong	gmtpyng	75	km t pyng	kim tae pyung
설렁탕	seor-reong-tang	srrngtng	75	sllngtng	seolleongtang
카피라이터	ka-pi-ra-i-teo	kprt	55	cpywrtr	copywriter
패션모델	pae-syeon-mo-der	psynmdr	53	fshn mdl	fashion model

Generate the Distractors

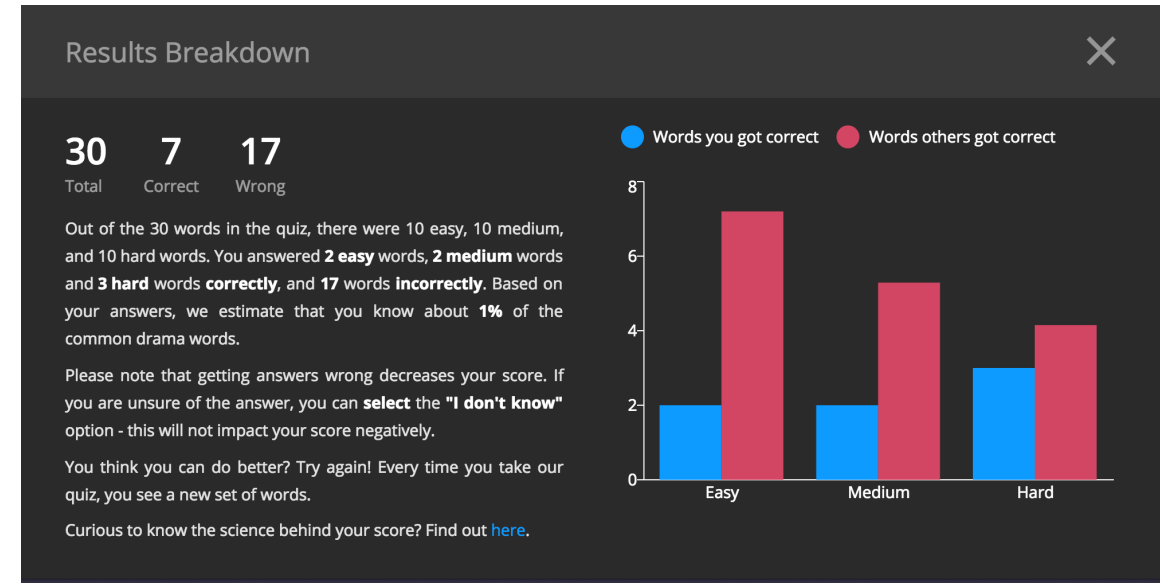
- Create a Word2Vec model on the source language
- For each word:
- **Distractor 1:** Select the top 5th to 20th closest words (*cosine*)
- **Distractor 2:** Use Distractor 1 as negative and question word as positive, select 1st to 20th closest word (*cosmul*)

Generate the Distractors



Splitting Word List into 3 Difficulty Levels

- **Word list after Step 1-5:**
 - Chinese: 14,000 common drama words
 - Korea : 8,000 common drama words
- **Using frequency as difficulty proxy:**
 - Split vocab into 3 difficulty bins;
 - ↑ frequency = simpler
 - Percentile ranges: (0-25], (25,75), [75,100]
 - Randomly select 10 words from each bin



**(A little) Smarter than
Absolute Score**

Your results



Hey,

Heol! (Omg!) You know about **11%** of the common drama words. It's a start, but you need to try harder to impress Jisung.



[SEE DETAILED RESULTS](#)

Share your result with friends. See who scores better!



How to Estimate learners' Vocab Size?

- Treat the no. of correct answers per level as observed mean of the test parameter from t-distribution.
- Confidence interval gives range of values parameterized by t-distribution such that observed mean is statistically probable.
- Find the lower & upper bound of confidence interval from one-tail t-test ($p < 0.1$)
- Multiply the lower/upper bound of CI by no. of words

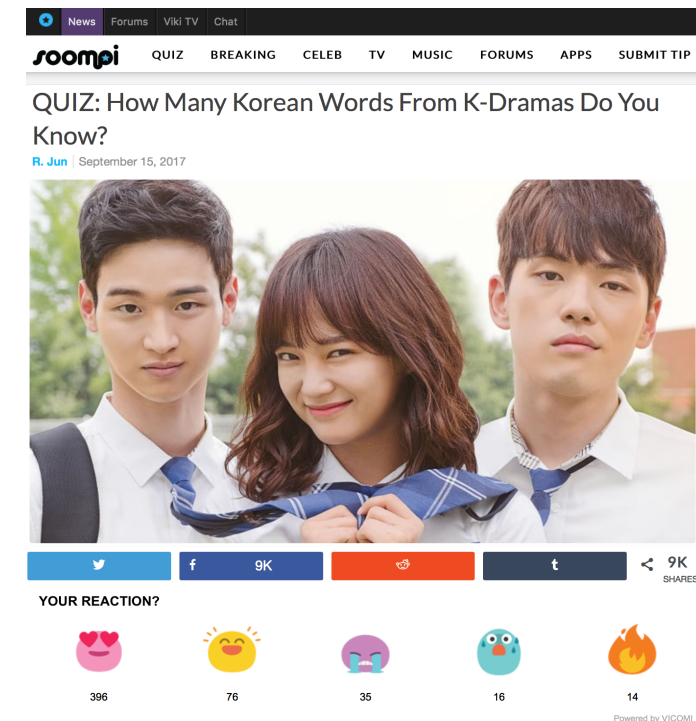
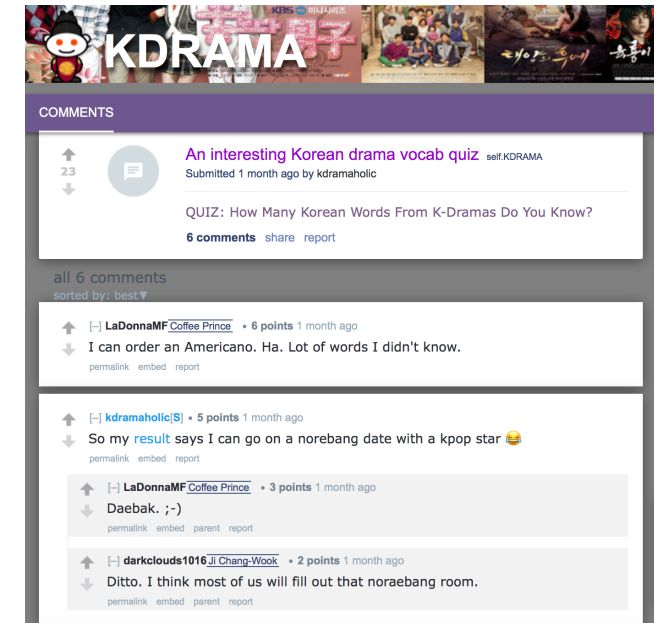
How to Estimate learners' Vocab Size?

- Let's say you answered 5 out of 10 correctly
- `input = [1, 1, 0, 1, 0, 0, 1, 0, 1, 0]`
- `stdev(input) = 0.527`
- `mean(input) = 5`
- One-tail t-test ($p < 0.1$), so cumulative frequency = 0.9
- Degree of freedom = `len(input) - 1 = 9`
- Critical value of t = `percent-point-function(0.9, 9) = 1.38`
- $CI = \text{mean} \pm (1.38 * 0.527) / \sqrt{\text{len(input)}} = 5 \pm 0.229 = (4.771, 5.229)$

Numbers and Graphs

Language Learners Like Quizzes!!

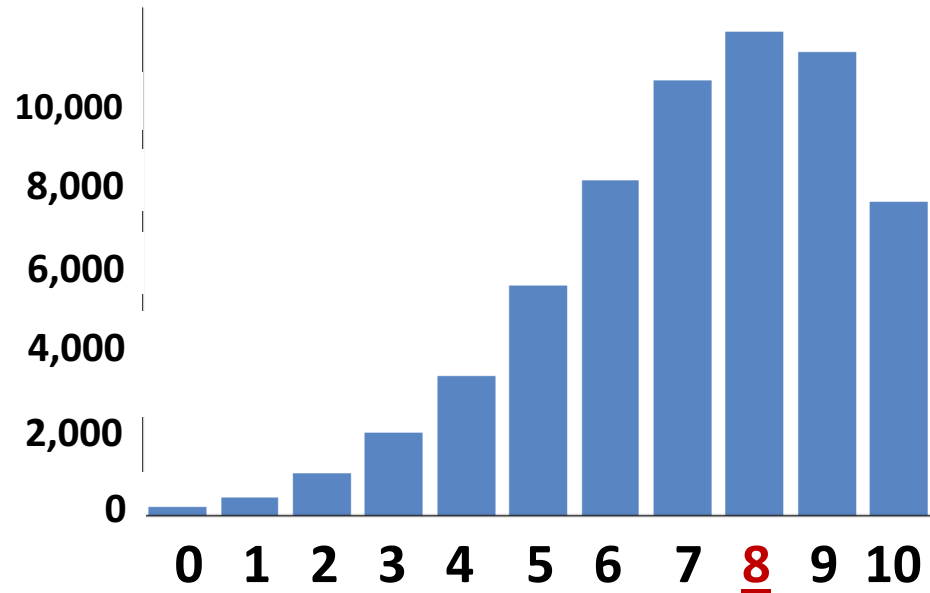
- 60,000+ quizzes taken in 2 weeks
- 35,000+ unique users completed quiz
- 16% of the users repeated quiz



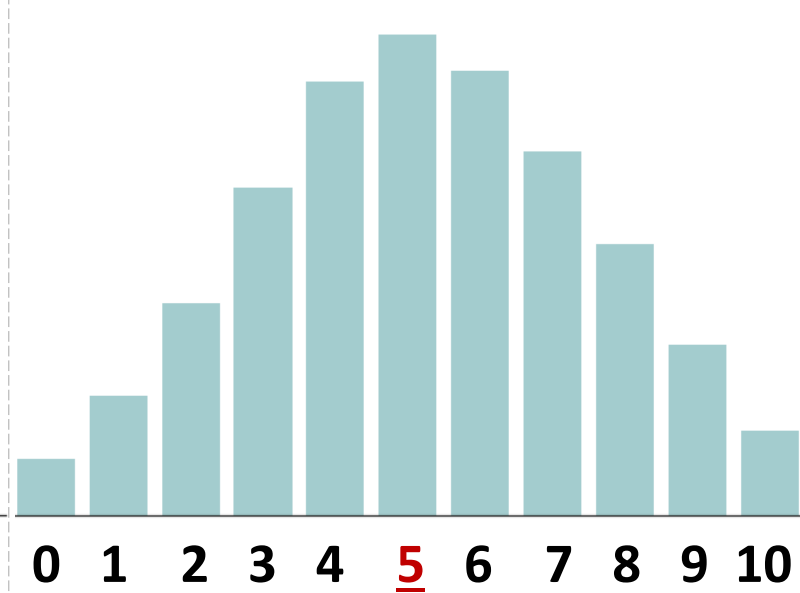
Is Word Frequency a Good Proxy to Word Difficulty?

Word Frequency is a Good Indicator for Difficulty Level

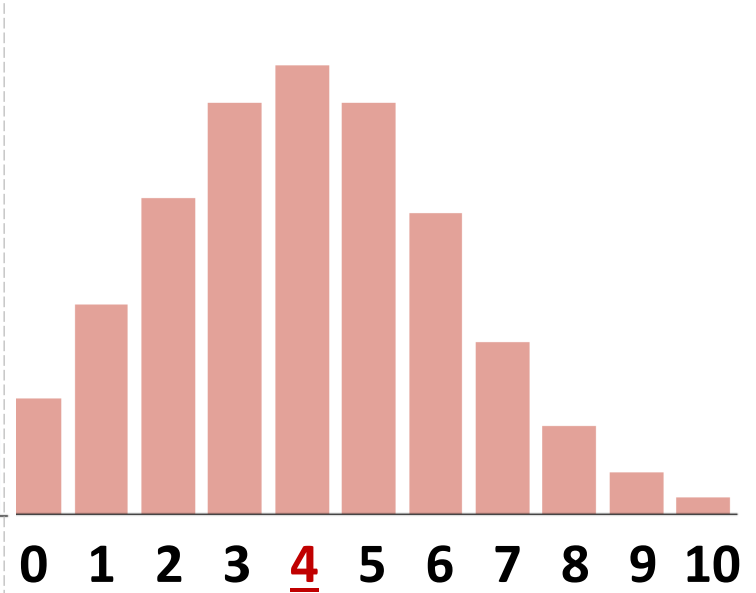
No. of **Easy Words**
Users got correct



No. of **Medium Words**
Users got correct



No. of **Hard Words**
Users got correct

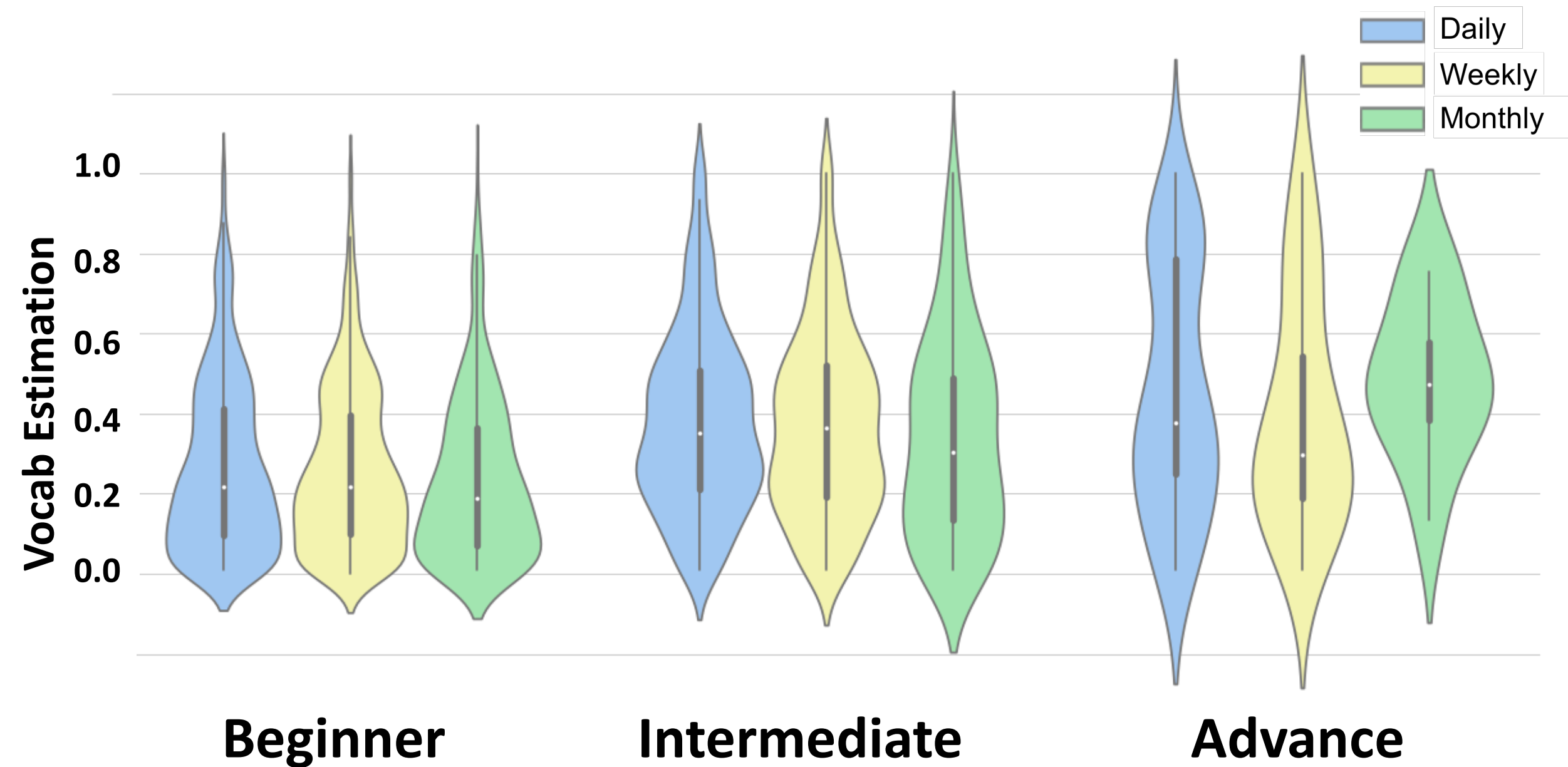


But it's not the Perfect Indicator...

Infrequent words can be simple

토마토 (tomato) 벨트 (belt) 한복 (hanbok) 아이스 (ice)
오케스트라 (orchestra) 아마추어 (amateur) 엘리베이터 (elevator)
화장품 (cosmetics) 섹스 (sex) 셀카 (selfie)
네티즌 (netizen) 불고기 (bulgogi) 샌드위치 (sandwich)
레드 (red) 액션 (action) 수박 (watermelon) 테이프 (tape)
코트 (coat) 시각 (time) 센스 (sense) 스키 (ski) 물고기 (fish) 주스 (juice)
코치 (coach) 로봇 (robot) 피자 (pizza) 라이터 (lighter)
리조트 (resort) 파트 (part) 파워 (power) 레슨 (lesson) 취한 (drunk)
순수 (pure) 밥그릇 (rice bowl) 스페셜 (special) 스탠바이 (standby)
패션 (fashion) 유머 (humor) 레시피 (recipe) 버터 (butter)
티켓 (ticket) 프로젝트 (project) 레몬 (lemon) 모니터 (monitor)
코끼리 (elephant) 사우나 (sauna) 섹시 (sexy)
메신저 (messenger) 테니스 (tennis) 아르바이트 (part-time job)

**Does watching K-drama more
often help you improve Korean?**



Conclusion

Conclusion

- **How to generate quiz automatically?**
- **Frequency is a good indicator of lexical complexity**
 - but we can do better
- **Advance learners benefit from watching dramas**
 - but not too much

Resources

Learn Mode: <https://viki.com>

Quiz: <https://languagequiz.viki.com>

Techblog: <https://techblog.rakuten.co.jp/2017/05/26/lang-quiz/>

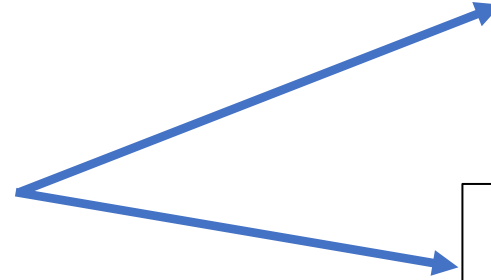
WordNets and CALL

WordNet for Language Learning

자격증:

{

Translations / Crosslingual WSD /
Annotated Corpus



자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

WordNet for Language Learning

자격증:

```
{  
  xling: {certificate: (sentid_1, sentid_2, ...),  
           license: (sentid_34, sentid_56, ...)  
        },  
}
```

자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

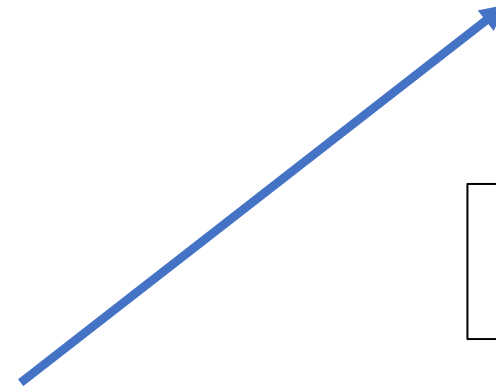
WordNet for Language Learning

자격증:

{

Translations / Crosslingual WSD /
Annotated Corpus

Lemma frequency information



자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

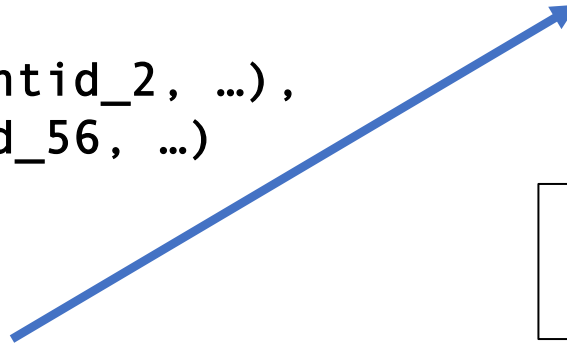
I don't know

WordNet for Language Learning

자격증:

```
{ xling: {certificate: (sentid_1, sentid_2, ...),  
          license: (sentid_34, sentid_56, ...)  
        },
```

```
counts: {sejong: (432, 0.03),  
         chosun: (2342, 0.008),  
         wiki:   (952, 0.045),  
         ...},
```



자격증 🗣️

[ja-gyeok-jeung]

Certificate

Document

Church

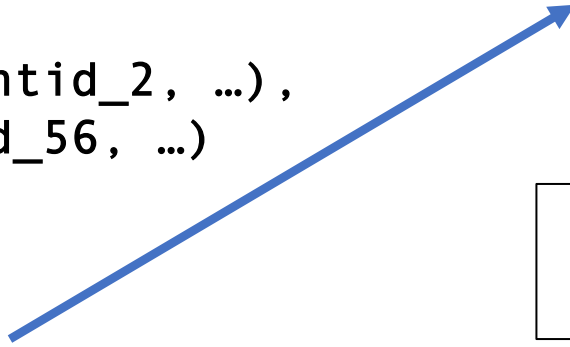
I don't know

WordNet for Language Learning

자격증:

```
{ xling: {certificate: (sentid_1, sentid_2, ...),  
          license: (sentid_34, sentid_56, ...)  
        },
```

```
counts: {DoTS: (432, 0.03),  
         LBS: (2342, 0.008),  
         BoF: (952, 0.045),  
         ... },
```



자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

WordNet for Language Learning

자격증:

- {
- Translations / Crosslingual WSD / Annotated Corpus
- Lemma frequency information
- Extra-modality: Video, Romanization, Speech

자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

WordNet for Language Learning

자격증:

```
{ xling: {certificate: (sentid_1, sentid_2, ...),  
          license: (sentid_34, sentid_56, ...)  
        },
```

```
counts: {sejong: (432, 0.03),  
         chosun: (2342, 0.008),  
         wiki:   (952, 0.045),  
         ...},
```

```
sound: {male:   (jageokjeung-m1.wav,  
                jageokjeung-m2.wav),  
        female: (jageokjeung-w1.wav,  
                jageokjeung-w2.wav)},
```

자격증 

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

WordNet for Language Learning

자격증:

```
{ xling: {certificate: (sentid_1, sentid_2, ...),
            license: (sentid_34, sentid_56, ...)
        },

  counts: {sejong: (432, 0.03),
            chosun: (2342, 0.008),
            wiki:   (952, 0.045),
            ...},

  sound: {male:   (jageokjeung-m1.wav,
                  jageokjeung-m2.wav),
          female: (jageokjeung-w1.wav,
                  jageokjeung-w2.wav),

  ontology: {synonym: {...}, related_to: {...}}

  forms:   {plural: {-을: (sentid_345, sentid_789, ...)},
            singular: {∅: (sentid_85, sentid_29, ...)} }

  collocations: {교사 자격증: (sent_523, ...), ... }  } } } } }
```

자격증

[ja-gyeok-jeung]

Certificate

Document

Church

I don't know

WordNet for Language Learning

- Can contextualization (frequency, example sentences, crosslingual mappings) be scaled to 100 M / 1B sentences at “user-able” quality?
- Moving towards Multi-Modality (Video, Speech, Romanization)
- Should Grammatical Knowledge (morphology, syntax, collocations) be Lexicalized?

What Makes a (Good) Language Learning Quiz Application?

What Makes a Language Learning Quiz Application?

- Generate Quiz Automatically

1 / 30

전화

jeon-hwa 

anticipation

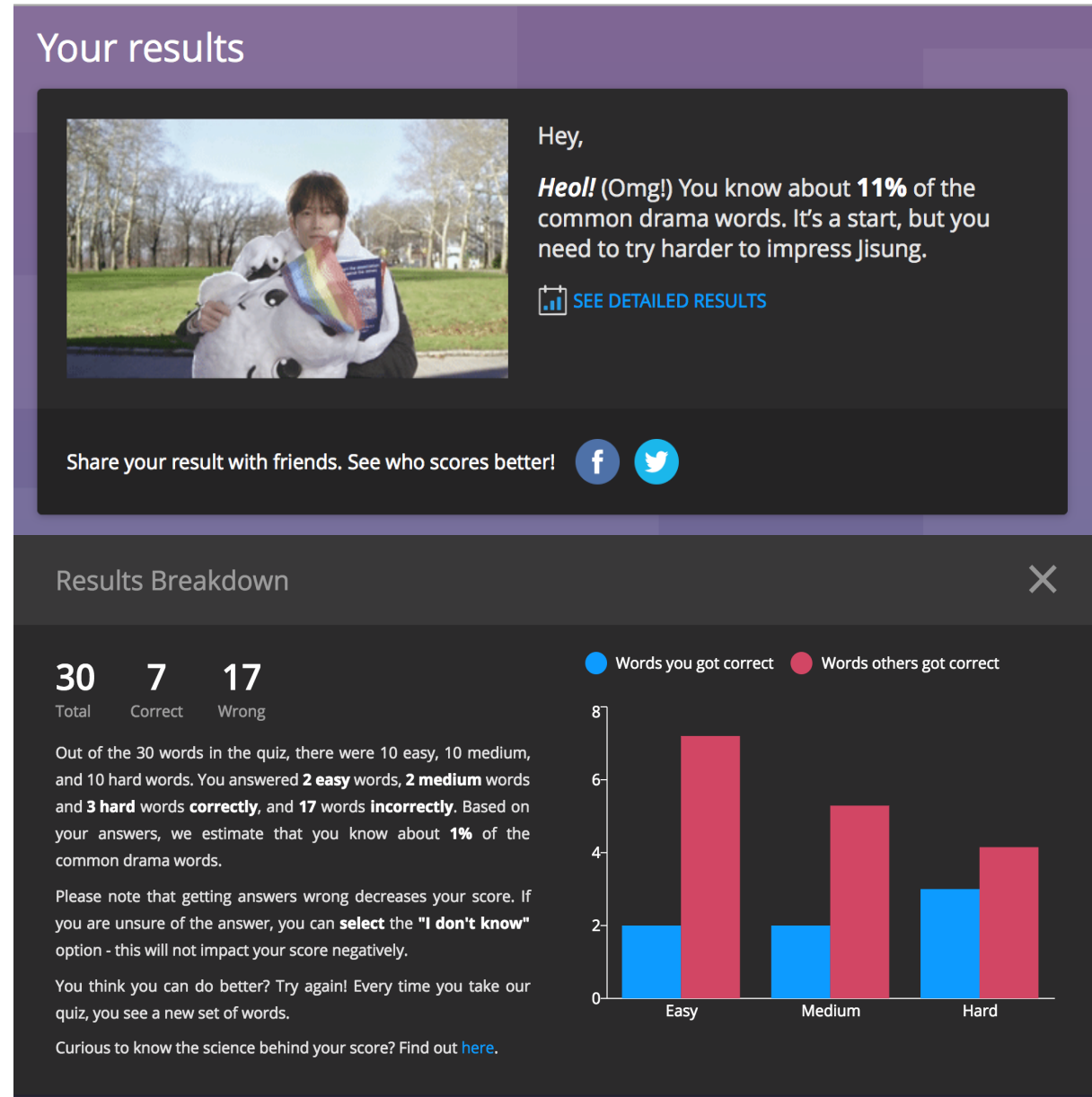
signal

telephone

I don't know >

What Makes a Language Learning Quiz Application?

- Generate Quiz Automatically
- Give Meaningful Results



What Makes a Language Learning Quiz Application?

- Generate Quiz Automatically
- Give Meaningful Results

Words you got wrong

Word	Romanization	Meaning
어제	eo-je	Yesterday
어머니	eo-meo-ni	Mother
남자	nam-ja	Man
상대	sang-dae	Opponent
얘기	yae-gi	Story

[SEE FULL LIST](#)

Words you got right

Word	Romanization	Meaning
대루	dae-ru	Counterwork
사랑	sa-rang	Love
패배	pae-bae	Defeat
연수	yeon-su	Soft water
참고	cam-go	Reference

[SEE FULL LIST](#)

What Makes a Language Learning Quiz Application?

- Generate Quiz Automatically
- Give Meaningful Results
- Reinforce Learning Experience

Watch and Learn

Here are some of the words you missed. With Learn mode, you can watch dramas and learn Korean at the same time. Check it out!

[WATCH AND LEARN](#)



민준 min-jun Min joon
This word appears in
EP 5: Oh My Venus



모순 mo-sun Contradiction
This word appears in
EP 8: I Hear Your Voice



실연 sir-yeon Disappointment in love
This word appears in
EP 1: It's Okay, That's Love



우동 u-dong Udon
This word appears in
EP 7: My Love From the Star

[SEE LESS](#)



Hey,

Heol! (Omg!) You know about **11%** of the common drama words. It's a start, but you need to try harder to impress Jisung.

 [SEE DETAILED RESULTS](#)

Share your result with friends. See who scores better!

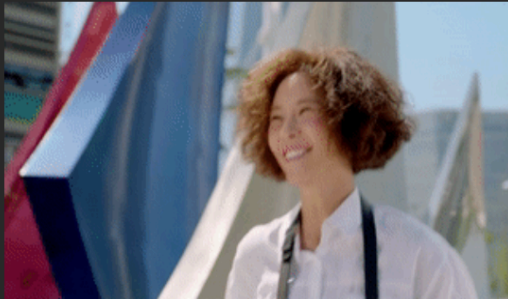


Hey,

Daebak! (Awesome!) You know about **4%** of the common drama words. You will survive a noraebang date with a pop star!

 [SEE DETAILED RESULTS](#)

Share your result with friends. See who scores better!



Hey,

Jjang! (Great!) You know about **4%** of the common drama words. You can probably order an Americano at a coffee shop in Hongdae.

 [SEE DETAILED RESULTS](#)

Share your result with friends. See who scores better!

