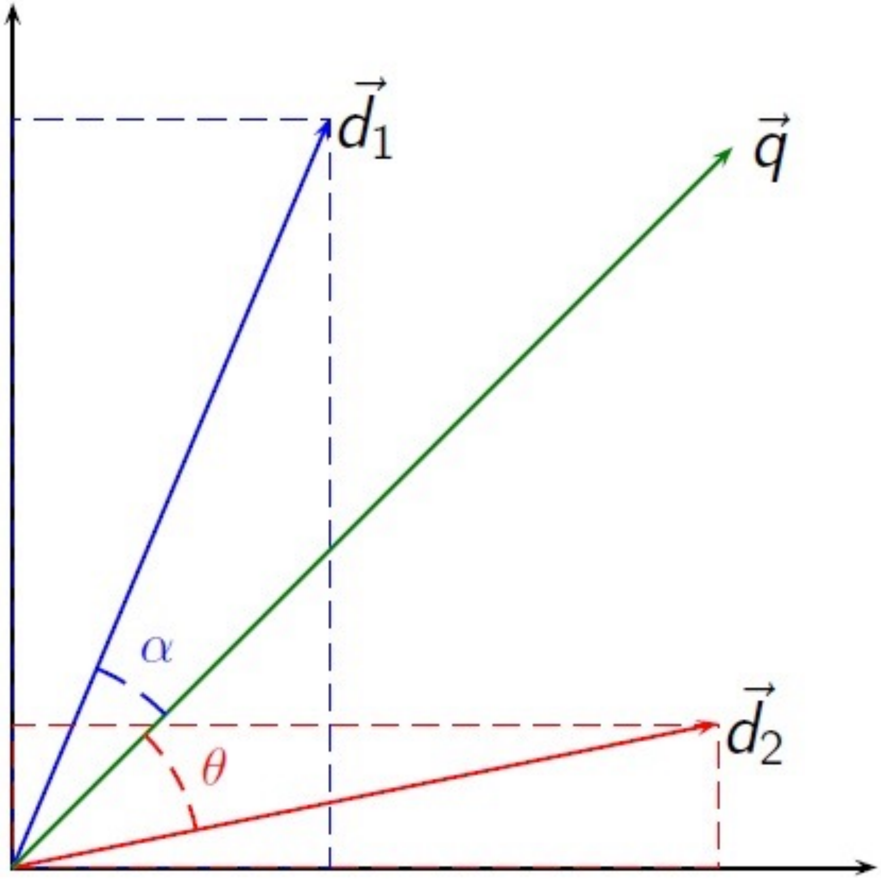
Lecture 23: Vector Semantics

Mark Hasegawa-Johnson

3/2023

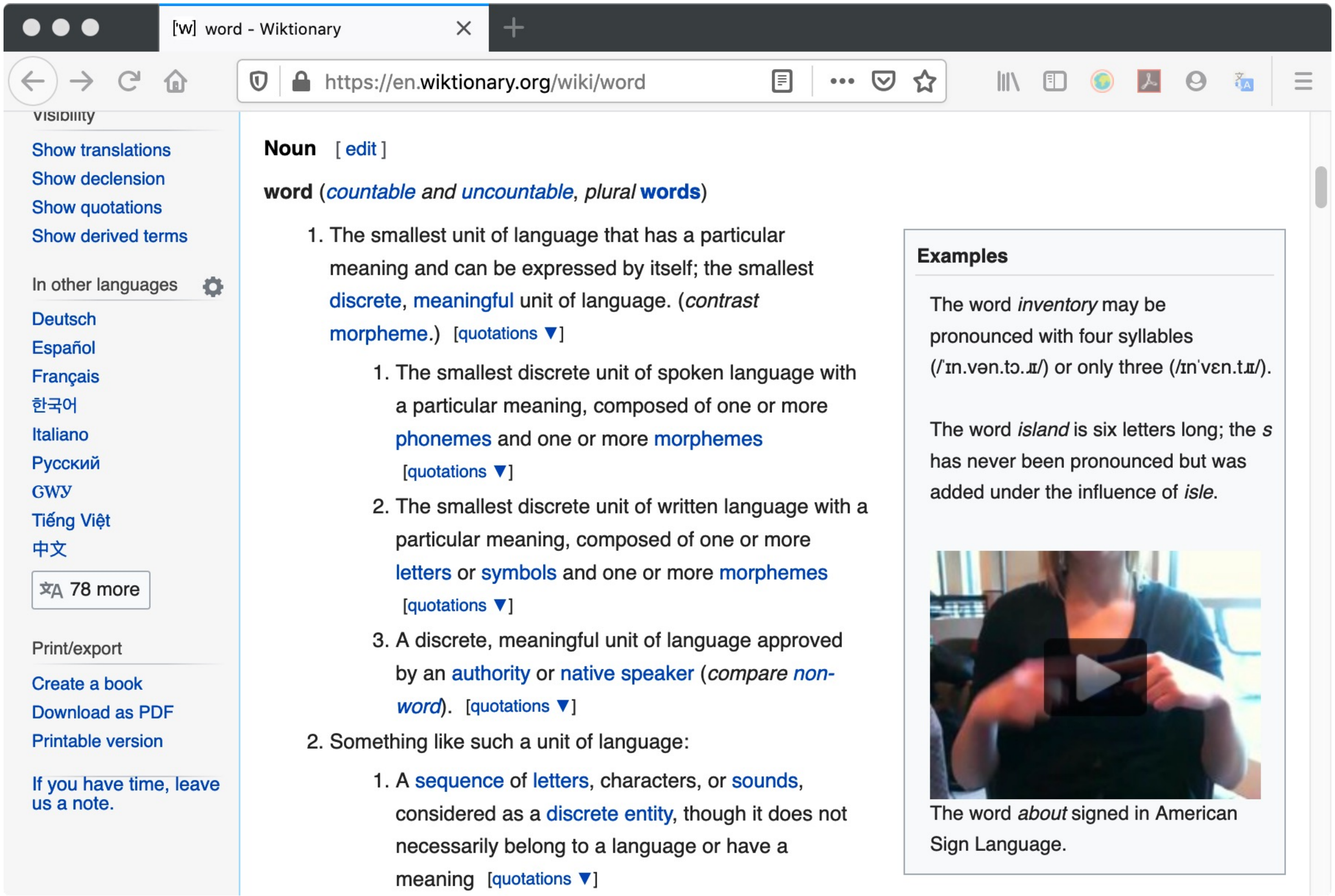


By Riclas - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=9076846

Outline

- What is a word? wordforms vs. lemmas vs. word senses
- What is meaning? synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

What is a word?



Noun [edit]

word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Examples

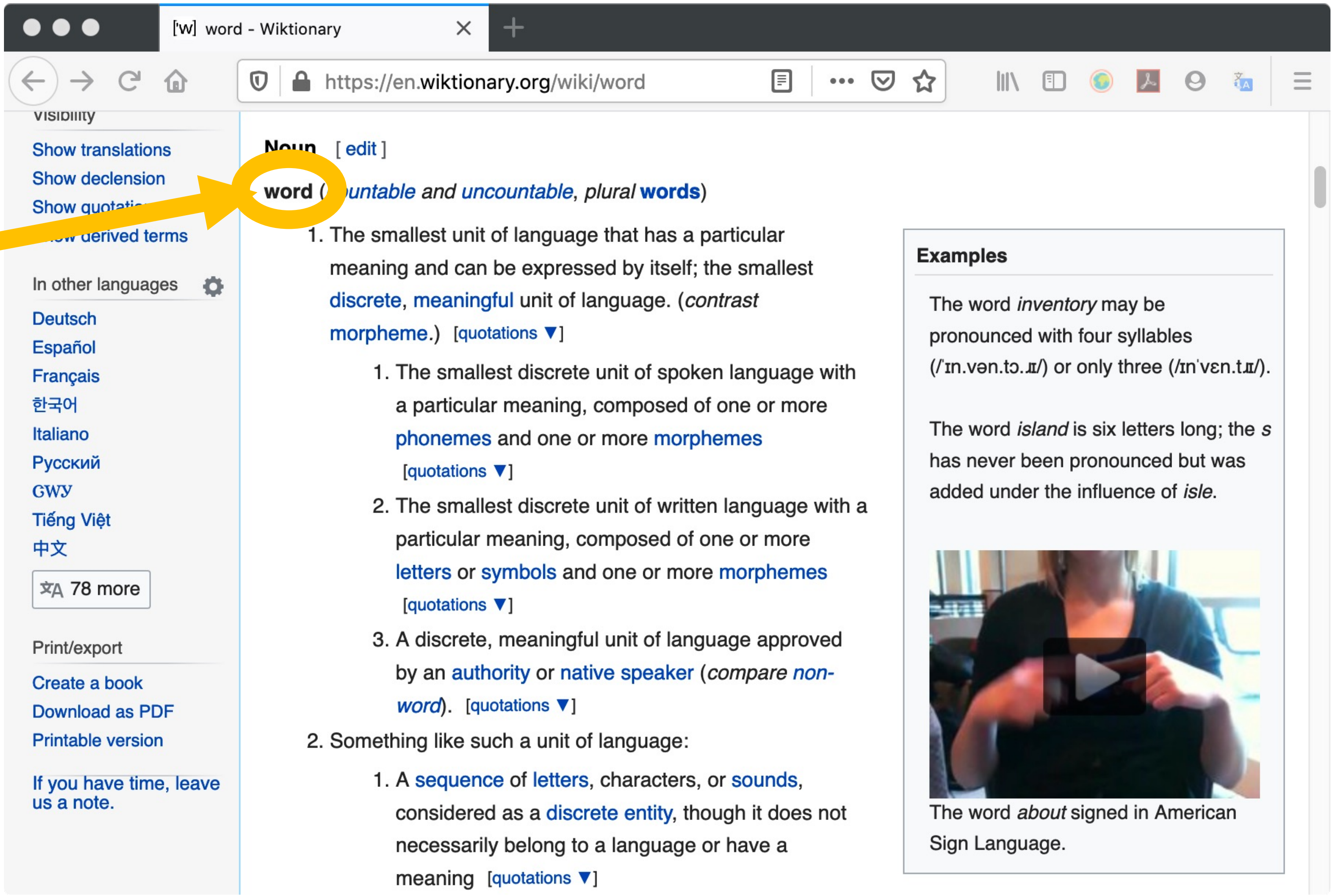
The word *inventory* may be pronounced with four syllables (/ˈɪn.vən.tɔ.ɹi/) or only three (/ɪnˈvɛn.tɹi/).

The word *island* is six letters long; the *s* has never been pronounced but was added under the influence of *isle*.

The word *about* signed in American Sign Language.

What is a word?

Is this a word?



Visibility

Show translations

Show declension

Show quotati

w derived terms

In other languages

Deutsch

Español

Français

한국어

Italiano

Русский

GWY

Tiếng Việt

中文

文A 78 more

Print/export

Create a book

Download as PDF

Printable version

If you have time, leave us a note.

Noun [edit]

word (countable and uncountable, plural **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast morpheme*.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Examples

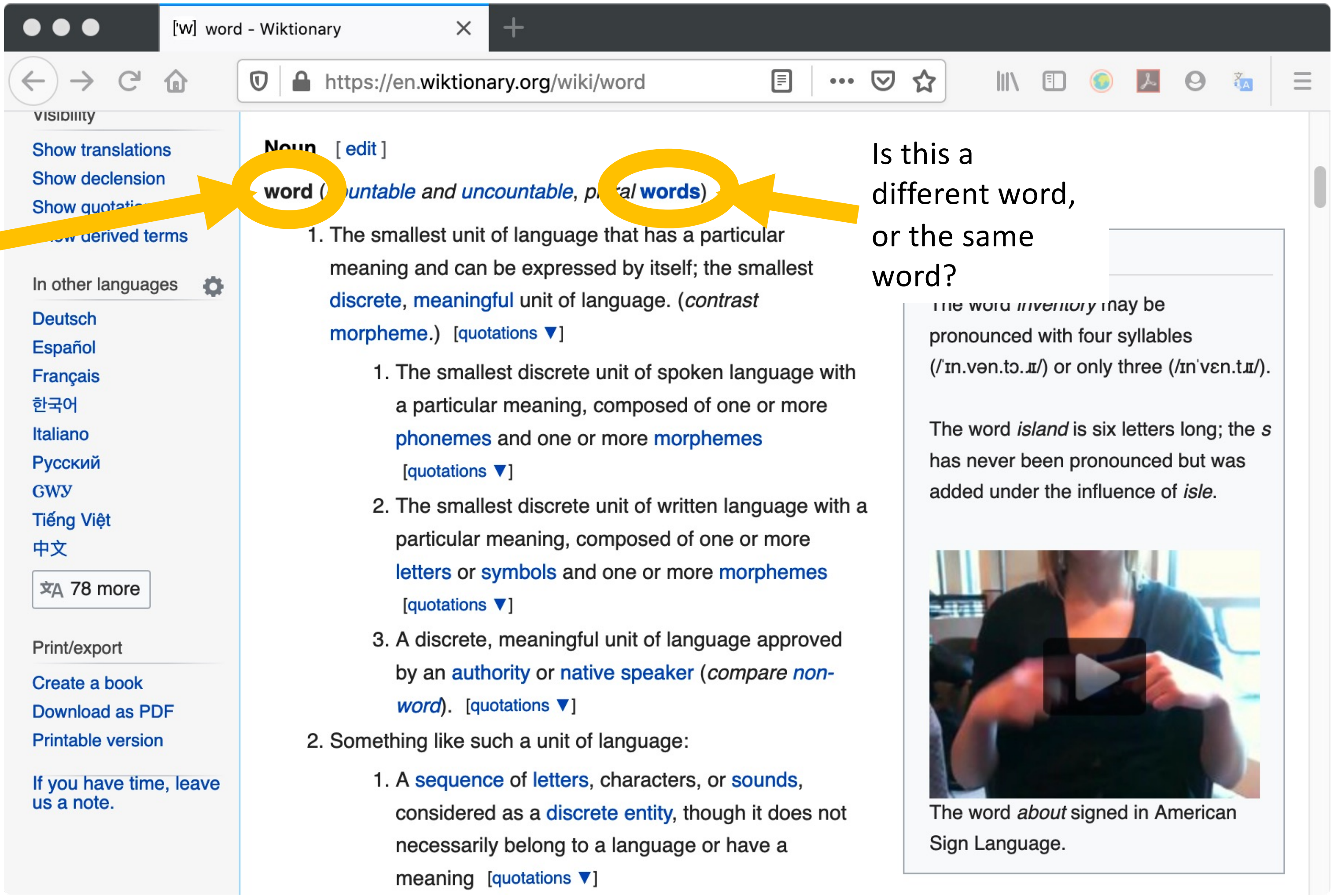
The word *inventory* may be pronounced with four syllables (/ˈɪn.vən.tɔ.ɹi/) or only three (/ɪnˈvɛn.tɹi/).

The word *island* is six letters long; the *s* has never been pronounced but was added under the influence of *isle*.

The word *about* signed in American Sign Language.

What is a word?

Is this a word?



Visibility

Show translations

Show declension

Show quotatio

ow derived terms

In other languages

Deutsch

Español

Français

한국어

Italiano

Русский

ᏣᎳᎩ

Tiếng Việt

中文

78 more

Print/export

Create a book

Download as PDF

Printable version

If you have time, leave us a note.

Noun [edit]

word (untable and uncountable, pl al **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast morpheme*.) [quotations ▼]
   1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
   2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
   3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
   1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

The word *inventory* may be pronounced with four syllables (/ˈɪn.vən.tɔ.ɹi/) or only three (/ɪnˈvɛn.tɹi/).

The word *island* is six letters long; the *s* has never been pronounced but was added under the influence of *isle*.

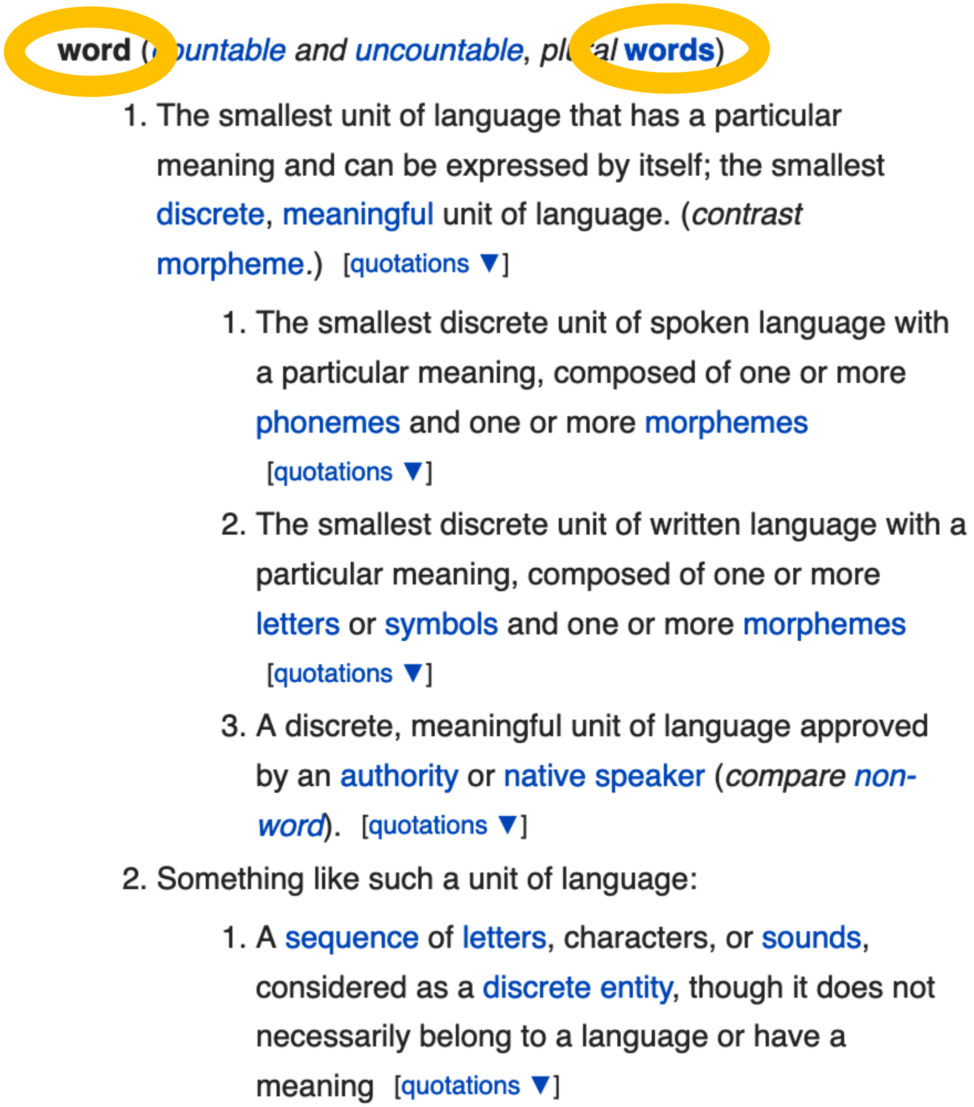
The word *about* signed in American Sign Language.

Is this a different word, or the same word?

Wordform

A wordform is a unique sequence of characters.

- Wordforms are much easier for computers to find than lemmas, therefore most automatic processing deals with wordforms.
- …however, we lose something. “dog” and “dogs” become completely unrelated – as unrelated as “dog” and “exaggerate.”



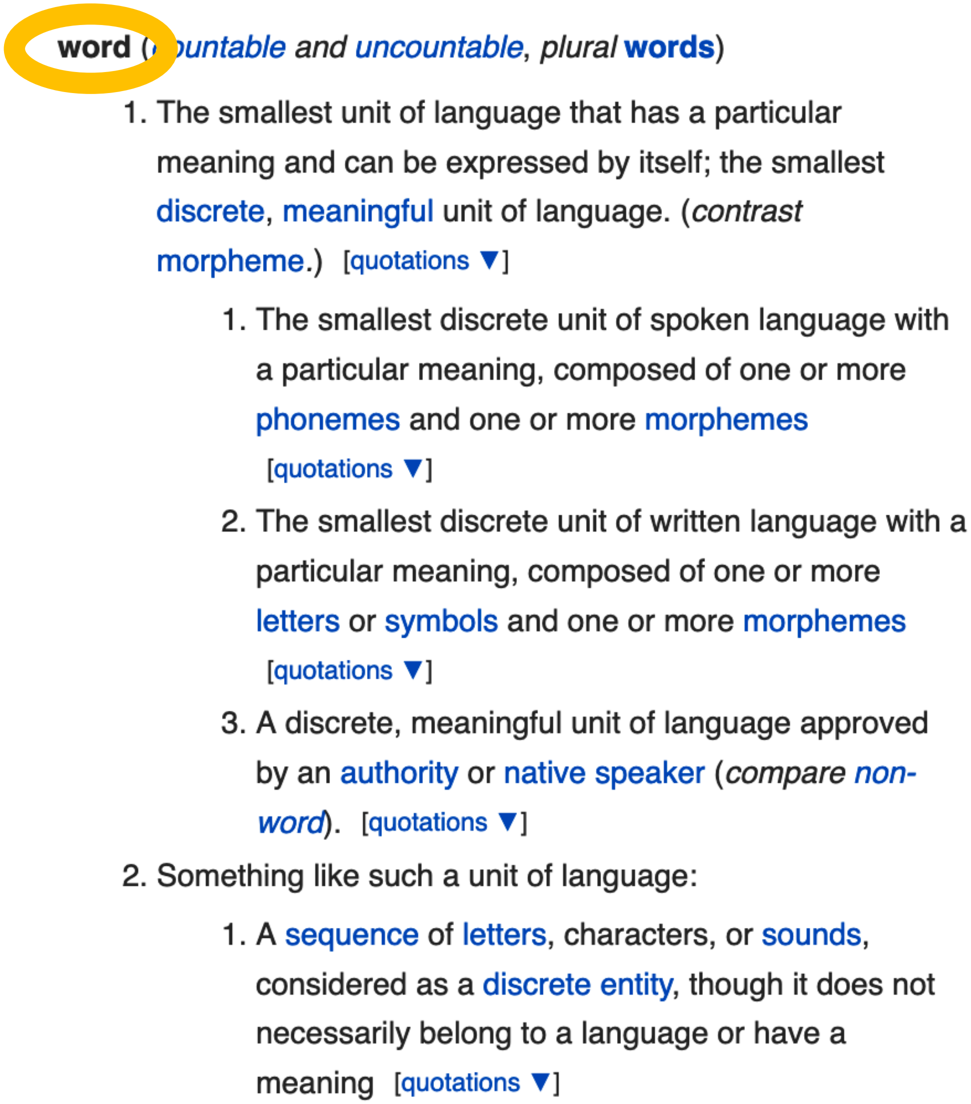
word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

Lemma

A lemma is what humans usually think of as a “word.” It is defined to be the form of the word that appears in a dictionary.

- Other wordforms that can be easily predicted from the lemma need not be listed.



word (*countable and uncountable*, *plural* **words**)

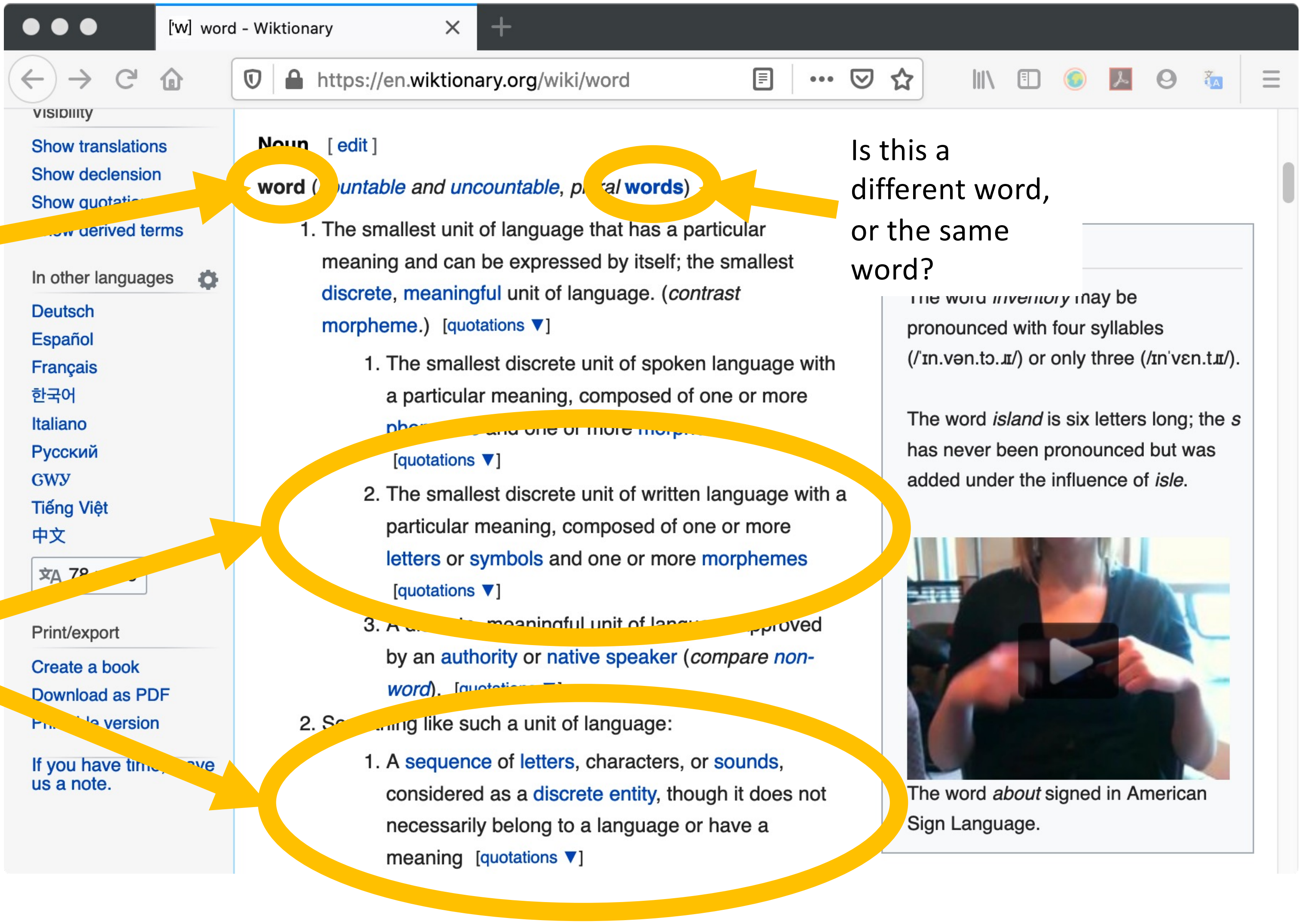
1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

What is a word?

Is this a word?

Are these the same word, or different words?

Is this a different word, or the same word?



Noun [edit]

word (*countable and uncountable*, *plural* **words**)

1. The smallest unit of language that has a particular meaning and can be expressed by itself; the smallest discrete, meaningful unit of language. (*contrast* morpheme.) [quotations ▼]
    1. The smallest discrete unit of spoken language with a particular meaning, composed of one or more phonemes and one or more morphemes [quotations ▼]
    2. The smallest discrete unit of written language with a particular meaning, composed of one or more letters or symbols and one or more morphemes [quotations ▼]
    3. A discrete, meaningful unit of language approved by an authority or native speaker (*compare non-word*). [quotations ▼]
2. Something like such a unit of language:
    1. A sequence of letters, characters, or sounds, considered as a discrete entity, though it does not necessarily belong to a language or have a meaning [quotations ▼]

The word *inventory* may be pronounced with four syllables (/ˈɪn.vən.tɔ.ɹ/) or only three (/ɪnˈvɛn.tɹ/).

The word *island* is six letters long; the *s* has never been pronounced but was added under the influence of *isle*.

The word *about* signed in American Sign Language.

Word sense

Often, a word has different meanings that are completely unrelated. We think of them as different words, that just happen to be spelled and pronounced the same way.

We say that these are different "senses" of the same word.

The Bank of England. By Diliff - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=40912212

The Bank of the Thames. By Diliff - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=3639626

Wordform, lemma, and word sense

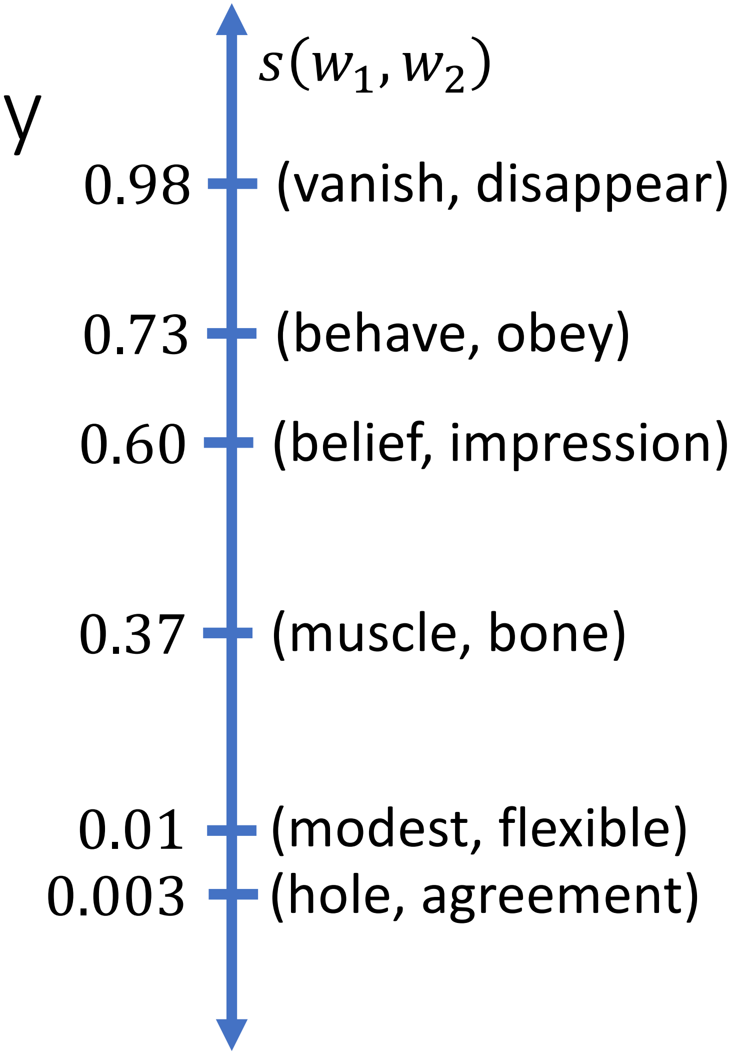
- wordform
  - easy for a computer to work with: just look for space-bounded sequences of characters
- lemma
  - This is what humans think of as a word. A cluster of wordforms whose spellings, pronunciations, and meanings can all be derived from one another by applying simple rules.
- word sense
  - A meaning so distinct from the other meanings of the word that it’s hard to consider them the same word.

Outline

- What is a word? wordforms vs. lemmas vs. word senses
- What is meaning? synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

Synonymy and similarity

- Words are “synonyms” if they have exactly the same meaning.
- No words ever have ***exactly*** the same meaning, so no two words are ever exactly synonyms.
- We prefer to talk about word similarity, $0 \leq s(w_1, w_2) \leq 1$
  - $s(w_1, w_2) = 1$: w_1 and w_2 are perfect synonyms. Never happens in practice, but sometimes close.
  - $s(w_1, w_2) = 0$: w_1 and w_2 are completely different.



$s(w_1, w_2)$

0.98 (vanish, disappear)

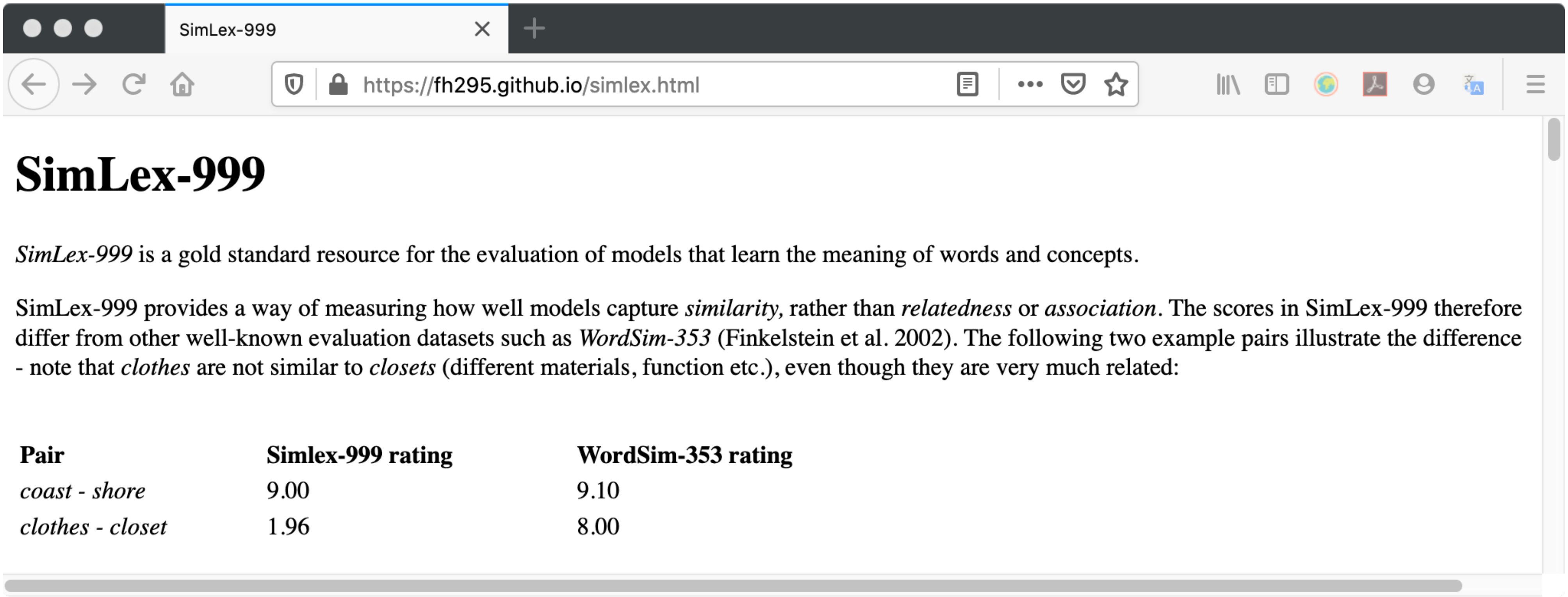
0.73 (behave, obey)

0.60 (belief, impression)

0.37 (muscle, bone)

0.01 (modest, flexible)

0.003 (hole, agreement)



SimLex-999

SimLex-999 is a gold standard resource for the evaluation of models that learn the meaning of words and concepts.

SimLex-999 provides a way of measuring how well models capture *similarity,* rather than *relatedness* or *association*. The scores in SimLex-999 therefore differ from other well-known evaluation datasets such as *WordSim-353* (Finkelstein et al. 2002). The following two example pairs illustrate the difference - note that *clothes* are not similar to *closets* (different materials, function etc.), even though they are very much related:

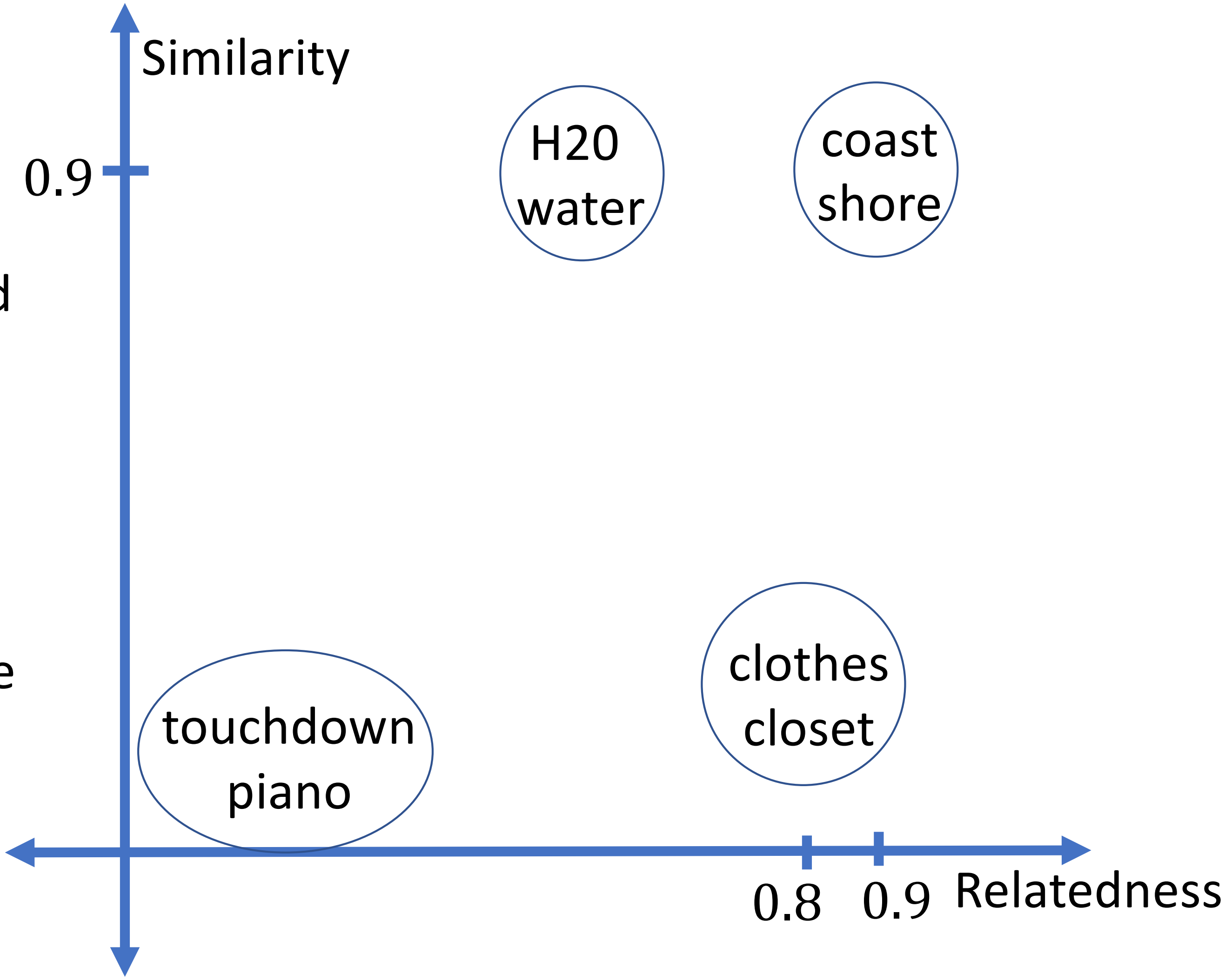
Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

- Algorithms that try to estimate the similarity of two wordforms can be tested on databases such as SimLex-999.
- Humans rated the similarity of each word pair on a 10-point scale.

Similarity vs. Relatedness

Similar: words can be used interchangeably in most contexts

Related: there is some connection between the two words, such that they tend to appear in the same documents.



Outline

- What is a word? wordforms vs. lemmas vs. word senses
- What is meaning? synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

Review: Naïve Bayes: the “Bag-of-words” model

We can estimate the likelihood of an e-mail by pretending that the e-mail is just a bag of words (order doesn’t matter).

With only a few thousand spam e-mails, we can get a pretty good estimate of these things:

- $P(W = \text{“hi”} | Y = \text{spam})$, $P(W = \text{“hi”} | Y = \text{ham})$
- $P(W = \text{“vitality”} | Y = \text{spam})$, $P(W = \text{“vitality”} | Y = \text{ham})$
- $P(W = \text{“production”} | Y = \text{spam})$, $P(W = \text{“production”} | Y = \text{ham})$

Then we can approximate $P(X|Y)$ by assuming that the words, W, are ***conditionally independent of one another given the category label***:

$$P(X = x | Y = y) \approx \prod_{i=1}^{n} P(W = w_i | Y = y)$$

Similarity: The Internet is the database

Similarity = words can be used interchangeably in most contexts

How do we measure that in practice?

Answer: extract examples of word w_1, +/- N words (N=2 or 3):

…hot, although iced **coffee** is a popular…

…indicate that moderate **coffee** consumption is benign…

…and of w_2:

…consumed as iced **tea**. Sweet tea is…

…national average of **tea** consumption in Ireland…

The words “iced” and “consumption” appear in both contexts, so we can conclude that $s(\text{coffea}, \text{tea}) > 0$. No other words are shared, so we can conclude $s(\text{coffee}, \text{tea}) < 1$.

skip-gram context probability

Consider the “…hot although iced **coffee** is a popular…”.

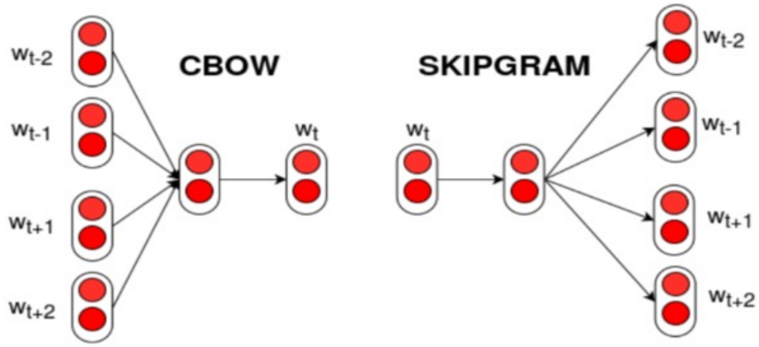
Define the target word to be w_0 =coffee.

Define the context words w_{-3} =hot, w_{-2} =although, …, w_3 =popular.

The skip-gram probability is a naïve Bayes model of the context:

$$p(w_{-3}, \dots, w_3 | w_0) = \prod_{\substack{i \neq 0 \\ i=-3}}^{3} p(w_i | w_0)$$

The skip-gram model



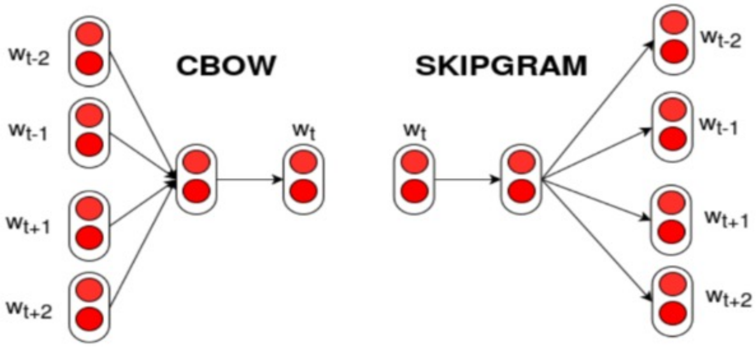
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_t|w_{t+j})$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_{t+j}|w_t)$$

- Skip-gram is a model of word meaning:
- The meaning of a word is defined to be the distribution of context words that it can predict.
- We find out which words w_t can predict by learning neural nets that predict its context words w_{t+j}:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=-c,j\neq 0}^{c} \ln P\left(w_{t+j}|w_t\right)$$

The “continuous bag of words” model (CBOW)



$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_t|w_{t+j})$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t)$$

- CBOW is a similar model of word meaning:
- The meaning of a word is defined to be the distribution of context words that predict it the best.
- We find out which words predict w_t by learning neural nets that predict w_t given its context words, w_{t+j}, for $-c \leq j \leq c$:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=-c, j\neq 0}^{c} \ln P\left(w_t|w_{t+j}\right)$$

“Probability,” for a NN, means softmax

- What does it mean that we train a neural net to compute $P\left(w_t|w_{t+j}\right)$?
- It’s a probability, so it must mean a softmax:

$$P\left(W_t = m|W_{t+j} = n\right) = \frac{\exp\left(e_{m,n}\right)}{\sum_{m\prime} \exp\left(e_{m\prime,n}\right)}$$

- But what are the inputs to the neural net? What is $e_{m,n}$?

Vector Semantics

- The simplest useful assumption is this: a word is a vector.

$$P\left(W_t = m | W_{t+j} = n\right) = \frac{\exp(v_m @ v_n)}{\sum_{m'} \exp(v_{m'} @ v_n)}$$

- …where v_m is a d-dimensional vector, $v_m = [v_{m,0}, \dots, v_{m,d-1}]$
- The only trainable parameters in this model are the word vectors!
- The dictionary, v , is a matrix, with as many rows as there are words in the vocabulary:

$$v = \begin{bmatrix} v_{\mathrm{a}} \\ \vdots \\ v_{\mathrm{zzz}} \end{bmatrix} = \begin{bmatrix} v_{\mathrm{a},0} & \cdots & v_{\mathrm{a},d-1} \\ \vdots & \ddots & \vdots \\ v_{\mathrm{zzz},0} & \cdots & v_{\mathrm{zzz},d-1} \end{bmatrix}$$

cosine similarity

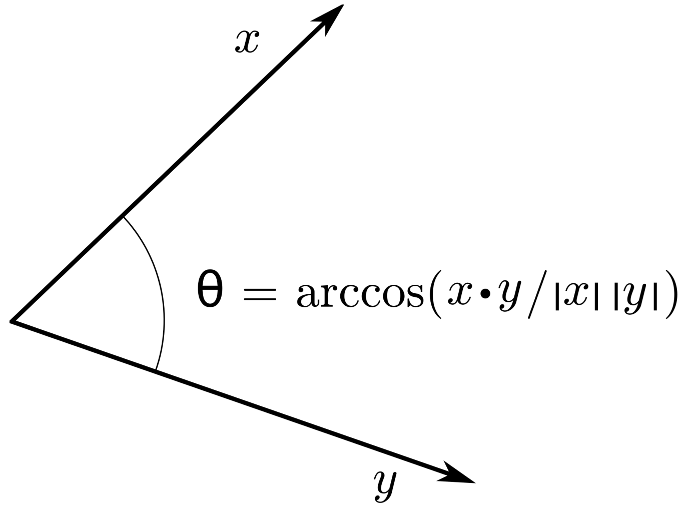
If words w_1 and w_2 are similar, w_1 is represented by vector $\vec{v}_1$, and w_2 by vector $\vec{v}_2$, then the angle between the two vectors should be small.

Angle between two vectors can be measured by their dot product:

$$\cos\theta = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1||\vec{v}_2|}$$

where

$$\vec{v}_1 \cdot \vec{v}_2 = \sum_{i=0}^{d-1} v_{1,i} v_{2,i}\,, \qquad |\vec{v}_1| = \sqrt{\sum_{i=0}^{d-1} v_{1,i}^2}$$



By BenFrantzDale at the English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=49972362

Vector Semantics

There are many ways to make this model more flexible. For example:

- Every word could have two different vectors: one (v_m) for when it's being predicted, one (c_n) for when it is predicting, thus $e_{m,n} = v_m @ c_n$.
- We could put a delay-weight matrix, w_j, in between the word vectors, thus $e_{m,j,n} = v_m @ w_j @ c_n$.
- We could even use an MLP to calculate the similarity, for example, $e_{m,n} = ReLU([v_m, v_n] @ w_0) @ w_1$.
- …but notice, all these methods are based on the idea of a matrix as a dictionary:

$$v = \begin{bmatrix} v_{\mathrm{a}} \\ \vdots \\ v_{\mathrm{zzz}} \end{bmatrix} = \begin{bmatrix} v_{\mathrm{a},0} & \cdots & v_{\mathrm{a},d-1} \\ \vdots & \ddots & \vdots \\ v_{\mathrm{zzz},0} & \cdots & v_{\mathrm{zzz},d-1} \end{bmatrix}$$

Vector Semantics

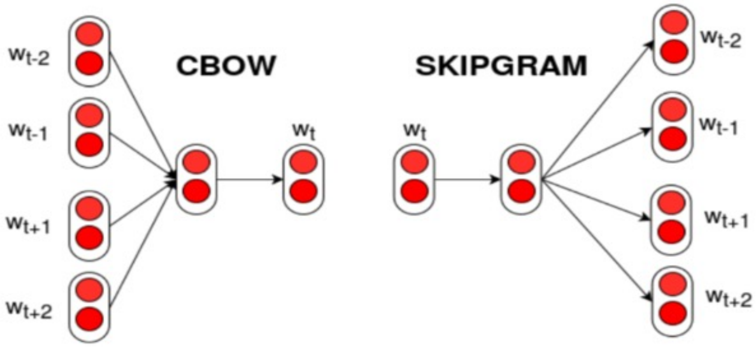
The CBOW probability is now:

$$P\left(W_t = m | W_{t+j} = n\right) = \frac{\exp(v_m @ v_n)}{\sum_{m'} \exp(v_{m'} @ v_n)}$$

Remember the derivative of a softmax:

$$\frac{\partial \mathrm{softmax}_m(e)}{\partial e_k} = \begin{cases} \mathrm{softmax}_m(e)(1 - \mathrm{softmax}_m(e)) & m = k \\ -\mathrm{softmax}_m(e)\mathrm{softmax}_k(e) & m \neq k \end{cases}$$

Training a CBOW model



$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_t|w_{t+j})$$

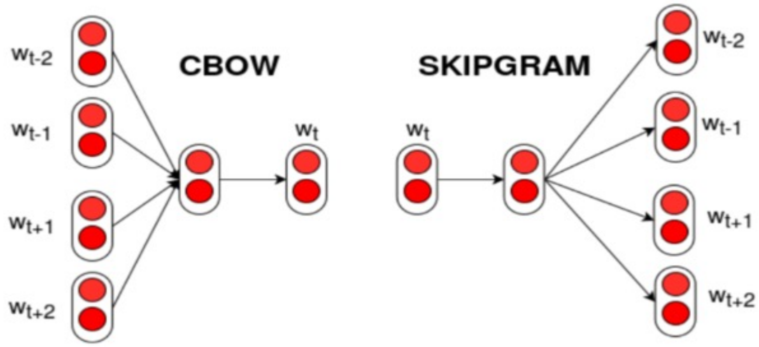
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_{t+j}|w_t)$$

In order to find the parameters, we use gradient descent:

$$\nabla_{v_m}\mathcal{L} = -\frac{1}{T}\sum_{t:w_t=m}\sum_{j=-c,j\neq 0}^{c}\nabla_{v_m}\ln P\big(W_t = m|w_{t+j}\big)$$

$$= -\frac{1}{T}\sum_{t:w_t=m}\sum_{j=-c,j\neq 0}^{c}\Big(1 - P\big(W_t = m|w_{t+j}\big)\Big)v_{w_{t+j}}$$

Training a CBOW model



$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_t|w_{t+j})$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t)$$

The CBOW model is trained by setting every vector equal to a weighted average of the words that occurred near it!

$$v_m \leftarrow v_m - \eta \nabla_{v_m} \mathcal{L}$$

$$v_m \leftarrow v_m + \frac{\eta}{T} \sum_{t:w_t=m} \sum_{j=-c, j\neq 0}^{c} \Big(1 - P\big(W_t = m | w_{t+j}\big)\Big)\, v_{w_{t+j}}$$

Try the quiz!

- Try the quiz: https://us.prairielearn.com/pl/course_instance/129874/assessment/2337428
- Vm = [1,0,0,0] + (eta/T)*((1-P(coffee|smells))*[0,0,0,1]+(1-P(coffee|hot))*[0,0,1,0])
- P(coffee|smells) = exp(vcoffee@vsmells)/sum(exp(vsmells@v))

= exp(0) / sum(exp(0)+exp(0)+ … + exp(0) + exp(vsmells@vsmells))

= 1 / ((N-1) + exp(1))

Outline

- What is a word? wordforms vs. lemmas vs. word senses
- Synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

Contrastive loss vs. Generative loss

- A generative loss is one like this:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=-c,j\neq 0}^{c}\ln\frac{\exp\left(v_{w_t}@v_{w_{t+j}}\right)}{\sum_{m'}\exp\left(v_{m'}@v_{w_{t+j}}\right)}$$

- Notice that this loss term compares each word, w_t, to every other word in the dictionary.
- Sometimes, generative training can take a very long time to converge.
- Sometimes, we get faster training using contrastive loss.

Contrastive loss

We train the neural network by listing, as positive examples, the words that occur in the context of "w =coffee," e.g.,

$$\mathcal{D}_+(w) = \{\text{hot}, \text{although}, \text{iced}, \text{moderate}, \text{hot}, \text{consumption}\}$$

Create a contrastive database by choosing the same number of words, at random, from among the words that never appeared in the context of "coffee:"

$$\mathcal{D}_-(w) = \{\text{aardvark}, \text{dog}, \text{gazebo}, \text{actor}, \text{precipitates}, \text{iceberg}\ \}$$

Training with contrastive loss

The coefficients $\vec{v}_i = \left[v_{i,0}, \dots, v_{i,d-1}\right]$ for each vector are then learned in order to maximize the log probability of the dataset:

$$\mathcal{L} = \ln p(\text{Data}) = \sum_{w \in \mathcal{V}} \ln p(\mathcal{D}_+(w)|w) + \sum_{w \in \mathcal{V}} \ln p(\mathcal{D}_-(w)|w)$$

$$= \sum_{w \in \mathcal{V}} \sum_{c \in \mathcal{D}_+(w)} \ln p(c|w) + \sum_{w \in \mathcal{V}} \sum_{c \in \mathcal{D}_-(w)} \ln\left(1 - p(c|w)\right)$$

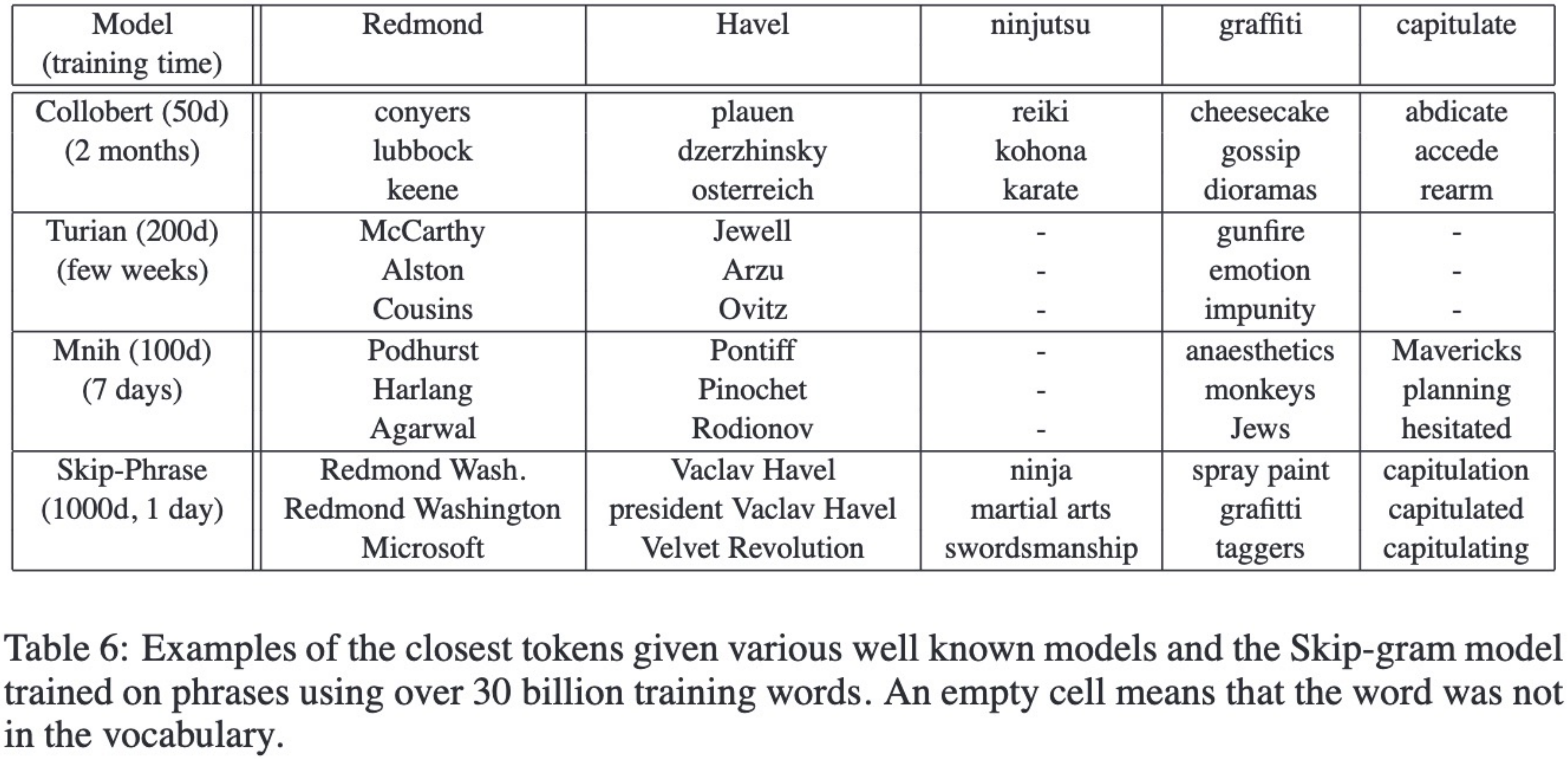
$$= \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_+(w)} \ln \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}} + \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_-(w)} \ln\left(1 - \frac{1}{1 + e^{-\vec{c}\cdot\vec{v}}}\right)$$

Outline

- What is a word? wordforms vs. lemmas vs. word senses
- Synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

Visualizations: Similarity

Mikolov et al. (2013) tested word2vec on SimLex-999, and had better results than previously published baselines. Here are some examples from their paper. Notice that not all of their ”similar words” are really similar – some are just related. I’ll talk more about that next time.

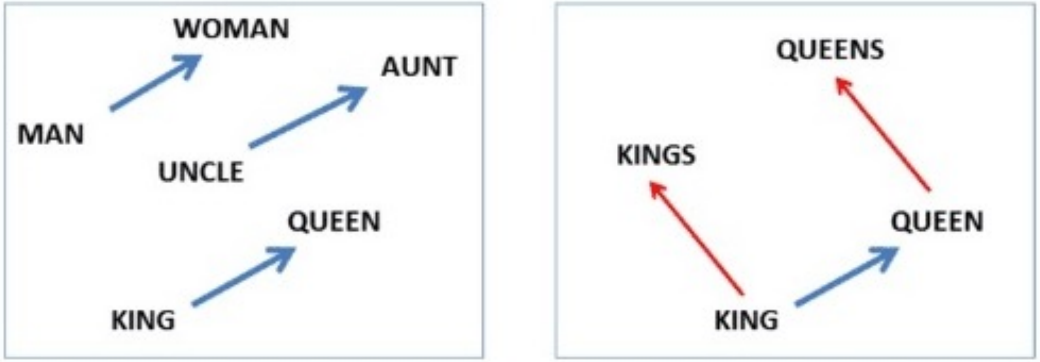


Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

Table 6: Examples of the closest tokens given various well known models and the Skip-gram model trained on phrases using over 30 billion training words. An empty cell means that the word was not in the vocabulary.

Visualizations: Relatedness

vec(“woman”) - vec(“man”) + vec(“king”) = vec(“queen”)



Christian S. Perone, "Voynich Manuscript: word vectors and t-SNE visualization of some patterns," in *Terra Incognita*, 16/01/2016, http://blog.christianperone.com/2016/01/voynich-manuscript-word-vectors-and-t-sne-visualization-of-some-patterns/.

Mikolov (2013) showed that word2vec captures similarity relationships among words. For example, the difference between the vectors for “woman” and “man” is roughly the same as the difference between the vectors for “queen” and “king.” Perone (2016) showed that this effect works differently depending on the training corpus: in his blog post, he looks at word relatedness in the 15th century Voynich manuscript.

Outline

- What is a word? wordforms vs. lemmas vs. word senses
- Synonymy, similarity, and relatedness
- Vector semantics: CBOW and skip-gram
- Generative training vs. Contrastive training
- Visualizations
- Bias

Learning biased analogies from data

- It’s useful that algorithms like word2vec learn appropriate analogies, like “Paris → France as Tokyo → Japan” and “kings → king as queens → queen.”
- Unfortunately, it also learns other analogies that were implied in the training corpus, but that are invalid analogies.
- The paper that first demonstrated that problem was named after one of the worst such discovered analogies:

“Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” Bolukbasi et al., 2016

Biased analogies

Bolukbasi et al. defined a “male-female” continuum by subtracting vec(female)-vec(male), vec(woman)-vec(man), and so on, then averaging these difference vectors.

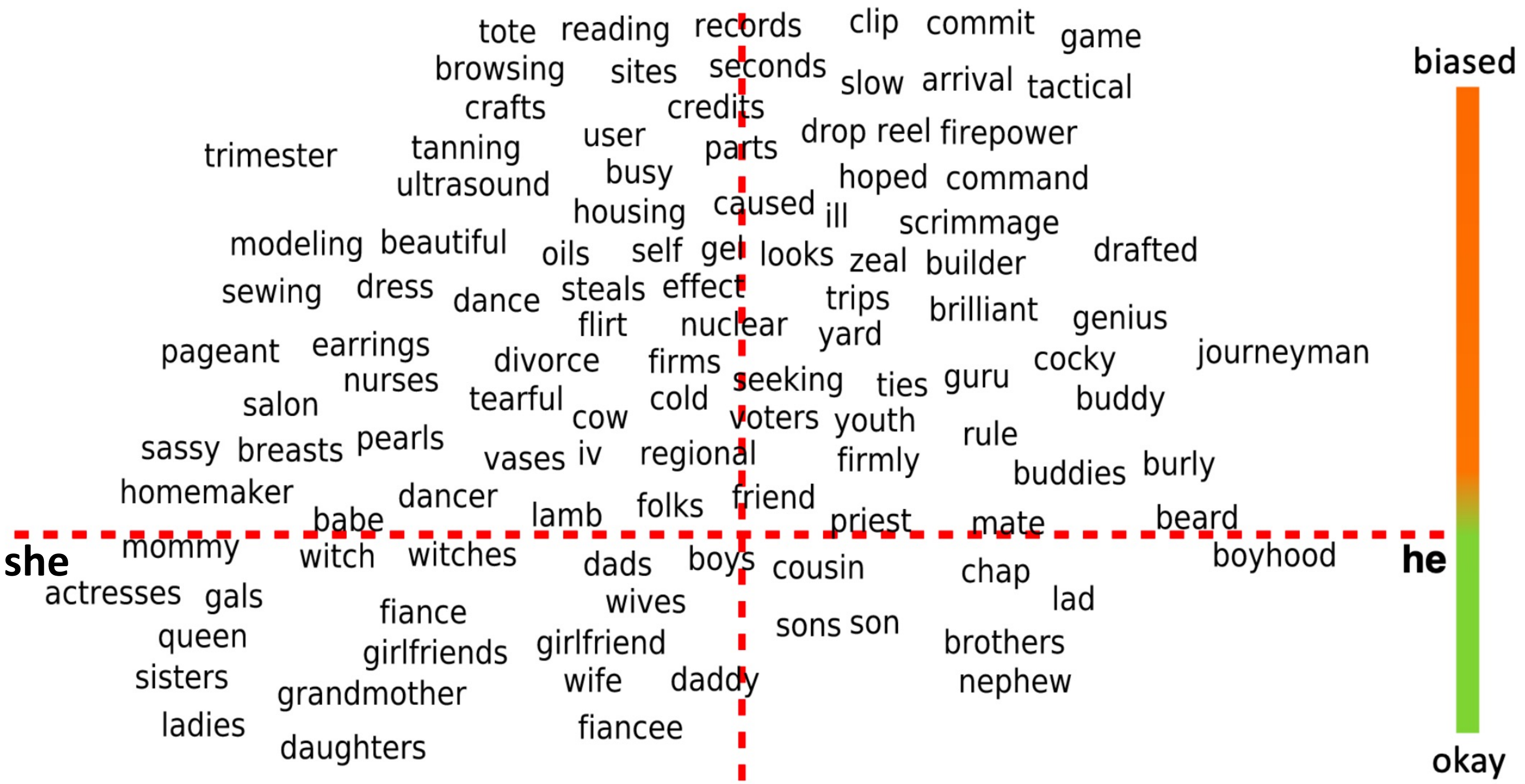
They then took all of the words whose dictionary definitions included gender-specific language (man, woman), and considered those to be the gender-specific words (words for which a gender difference is appropriate).

All other words were considered gender-neutral (any difference on the male-female dimension is inappropriate).

The result is a second dimension: the appropriateness of a gender bias.

The Male-Female vs. Neutral-Specific Space

Here’s the resulting 2D space, from Bolukbasi et al., 2016:



Outline

- What is a word? Lemmas, wordforms, and word sense
- Synonymy, similarity, and relatedness
- Word2vec: maximize

$$\mathcal{L} = \sum_{\vec{v} \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_+(w)} \ln \frac{1}{1 + e^{-\vec{c} \cdot \vec{v}}} + \sum_{w \in \mathcal{V}} \sum_{\vec{c} \in \mathcal{D}_-(w)} \ln \frac{1}{1 + e^{\vec{c} \cdot \vec{v}}}$$

- Visualizations
  - Similarity: list the K-nearest neighbors, show that they are similar
  - Relatedness: analogies are shown as directions in the vector space!
- Bias
  - Bias can be reduced by learning a direction that should not depend on the female-male axis, and then squashing the female-male axis to zero for words that should be gender-neutral.