ASSOCIATION MEASURES

Association measures refer to the relationship between a node word and its collocates, and these collocates provide meaning about the meaning and usage of the node word. For example, near the node word *nutrient* we might find the collocates {*vitamin, diet, healthy*}, and near the node word *telescope* we might find { *observatory, mirror, astronomer* }.

Researchers have developed several statistics to predict what collocates might co-occur with a given node word. For example, at Sketch Engine there are seven different statistics, which correspond to the rows #4-10 in the table below (see also Evert 2008 for a discussion of the different measures). (And see the page dealing with topics – words that co-occur *anywhere in the text*).

	telescope	**fresco**[1]	**kombucha**	**infrared**	**recyclable**	**evoke**	**excavate**	**consciously**
English-Corpora.org								
[1] Freq+MI (MI = 3.5)	space, radio, sky, optical, binoculars, observatory, observe, powerful, mirror, astronomer, lens, mount	painting, paint, wall, ceiling, chapel, church, decorate, depict, century, mosaic, beautiful	tea, kefir, brew, brewing, drink, batch, fermented, bottle, ferment, sauerkraut, drink, culture	light, camera, sauna, sensor, radiation, heat, visible, IR, spectrum, wavelength, LED, remote	material, plastic, packaging, waste, item, fully, paper, bottle, reusable, recycled, biodegradable, bag	emotion, feeling, memory, image, sense, response, emotional, spirit, potential, reaction, mood, nostalgia	site, archaeologist, soil, hole, rock, tunnel, tomb, trench, ancient, pit, ruin, burial	unconsciously, aware, choose, subconsciously, deliberately, relax, breathe, perceive, acknowledge, intentionally, reject, actively
[2] Freq+MI (MI = 6)	optical, binoculars, observatory, astronomer, infrared, eyepiece, ground-based, aperture, kepler, microscope, astronomical, focal	painting, paint, ceiling, chapel, decorate, depict, mosaic, renaissance, adorn, sculpture, Sistine	tea, kefir, brew, brewing, batch, fermented, ferment, sauerkraut, vinegar, kimchi, homemade, brew	sauna, sensor, radiation, visible, IR, spectrum, wavelength, LED, ultraviolet, thermometer, heater, laser	material, plastic, packaging, waste, reusable, recycled, biodegradable, compostable, recycle, aluminum, recycling, renewable	emotion, feeling, nostalgia, imagery, sympathy, auditory, nostalgic, vividly, bygone, powerfully, visceral, somatosensory	archaeologist, tunnel, tomb, trench, ruin, nest, burial, archaeological, fossil, grave, remains, artifact	unconsciously, aware, subconsciously, deliberately, purposefully, uncouple, uncoupled, mindfully, rationally, willfully, co-create, conscientiously
[3] MI(+freq)	hubble, keck, refractor, ground-based, kepler, earth-based, celestron, gamma-ray, space-based, focuser, spectrograph, Newtonian	Sistine, crumbled, mosaic, Byzantine, chapel, adorn, mural, renaissance, depict, ceiling, decorate	kefir, kimchi, sauerkraut, fermented, unflavored, fizzy, ferment, miso, brew, brewing, store-bought, probiotic	thermography, wide-field, electro-optical, spectroscopy, sauna, illuminator, spectrometer, ultraviolet, space-based, flir, imager, IR	compostable, biodegradable, reusable, polypropylene, recycled, curbside, nontoxic, packaging, polyethylene, carton, eco-friendly, infinitely, cardboard	somatosensory, nostalgia, emotion, bygone, auditory, vividly, nostalgic, powerfully, imagery, visceral, grandeur, feeling	archaeologist, archeologist, burrow, archeological, trench, archaeological, tomb, digger, buried, skeleton, remains	unconsciously, subconsciously, uncouple, uncoupled, deliberately, purposefully, aware, subconscious, incompetent, intentionally, purposely, disregard
Sketch Engine: good results								
[4] LogDice	hubble, observatory, telescope, spitzer, binoculars, ground-based, webb, nasa, infrared, subaru	al, giotto, sistine, dining, dine, fresco, michelangelo, mosaic, jacque	kombucha, kefir, scoby, kimchi, health-ade, sauerkraut, ferment, humm, probiotic, kevita	spectroscopy, infrared, IR, sauna, wavelength, fourier, radiation, spectrometer, wide-field, thermometer	compostable, reusable, biodegradable, recycled, packaging, non-recyclable, recyclable, cardboard, plastic, non-toxic	emotion, nostalgia, feeling, auditory, sympathy, imagery, vividly, memory, mood, potential	archaeologist, trench, tomb, remains, archaeological, archeologist, burrow, mound, pit, tunnel	unconsciously, subconsciously, deliberately, aware, intentionally, consciously, sub-consciously, uncouple, purposefully, nonconsciously
[5] MI.log-f	hubble, spitzer, ground-based, canada-france-hawaii, cherenkov, keck, binoculars, observatory, hobby-eberly, telescope	al, dining, giotto, sistine, dine, brancacci, buon, paint, mosaic	kombucha, kefir, scoby, health-ade, ferment, tea, kimchi, sauerkraut, brew, kevita	wide-field, spectroscopy, thermography, ftir, fourier, sauna, infrared, IR, spectrometer, radiation	compostable, reusable, biodegradable, packaging, material, recycled, plastic, non-recyclable, waste, cardboard	emotion, somatosensory, feeling, nostalgia, memory, auditory, response, myogenic, otoacoustic, somatosensory	archaeologist, trench, tomb, archeologist, remains, archaeological, site, burrow, tunnel, mound	unconsciously, subconsciously, sub-consciously, nonconsciously, aware, deliberately, choose, uncouple, intentionally, consciously
[6] Minimum sensitivity	hubble, telescope, observatory, spitzer, webb, nasa, infrared, binoculars, ground-based, astronomer	al, fresco, mosaic, giotto, michelangelo, dine, sistine, dining, raphael	kombucha, kefir, kimchi, sauerkraut, scoby, ferment, health-ade, probiotic, humm, miso	spectroscopy, infrared, IR, wavelength, sauna, fourier, thermometer, radiation, spectrometer, ultraviolet	compostable, biodegradable, reusable, recycled, packaging, recyclable, non-toxic, polypropylene, cardboard, non-recyclable	emotion, nostalgia, sympathy, auditory, imagery, feeling, vividly, mood, empathy, laughter	archaeologist, trench, remains, tomb, burrow, archaeological, mound, archeologist, fossil, excavate	unconsciously, subconsciously, deliberately, consciously, aware, intentionally, disregard, willingly, imitate, sub-consciously

[1] Tests #1-3 (our approach), #4-6 and #8 (Sketch Engine) all include the collocate *queso*, which is from the Spanish phrase *queso fresco*. We have removed that one word here.

	telescope	fresco[1]	kombucha	infrared	recyclable	evoke	excavate	consciously
Sketch Engine: Too specific / lower frequency words (and then sometimes changing to high frequency / not useful words for MI3)								
[7] MI-score	five-hundred-meter, siderostat-type, canada-france-hawaii, metrewave, hobby-eberly, meterwave, 102ed, mcmath-pierce, 64-m, thirty-meter	murideo, six-a, six-g, anossov, plaincourault, brancacci, rottmayr, guidoriccio, bull-leaping, schifanoia	health-ade, kevita, scoby, konnection, kombucha, kefir, booch, cold-brew, bucha, kvass	reflectography, bds2003, fourier-transform, reflection-absorption, visible-near, cedip, reflectogram, wide-field, raytemp, slitless	amlite, repulpable, compostable, mono-material, technopolymer, non-recyclable, pvc-free, compostable, bio-degradable, biodegradable	ssveps, sseps, somatosensory, ssvep, bio-potential, teoaes, baep, vemps, otoacoustic, teoae	dibbits, pfas-contaminated, sondage, mellaart, occaneechi, mawangdui, archaeologically, fasciole, mallowan, forrestfield	nonconsciously, sub-consciously, unconsciously, subconsciously, volitionally, uncouple, mindfully, subliminally, co-create, conscientiously
[8] MI3	hubble, space, the, spitzer, telescope, observatory, ground-based, ., radio, binoculars	al, dining, paint, the, dine, painting, giotto, sistine, .	kombucha, kefir, tea, health-ade, scoby, ferment, kimchi, sauerkraut, kevita, brew	spectroscopy, wide-field, radiation, light, sensor, infrared, camera, IR, wavelength, sauna	compostable, material, reusable, packaging, biodegradable, waste, plastic, recycled, be, and	emotion, the, of, feeling, a, ,, memory, and, response, that	archaeologist, be, the, site, ., ,, and, in, trench, of	unconsciously, subconsciously, aware, to, ,, be, or, not, and, .
Sketch Engine: Too general / high frequency words								
[9] Log likelihood	the, hubble, space, ., a, ,, and, radio, large, of	al, the, ., of, ,, dining, and, in, paint, painting	tea, kombucha, ,, ., ferment, and, be, kefir, a, the	light, the, and, ., ,, camera, radiation, sensor, spectroscopy, (	material, be, and, packaging, ., ,, %, reusable, waste, compostable	the, of, a, ,, and, that, to, emotion, ., feeling	be, the, site, ., ,, and, in, of, have, to	unconsciously, to, be, ,, or, not, and, ., the, aware
[10] T-score	the, ., ,, a, and, be, of, space, in, with	the, ., ,, of, and, al, in, a, be, with	,, ., and, be, the, a, of, tea, in, make	the, ., ,, and, a, of, (, be,), light	be, and, ., ,, material, of, the, [number]-m, %, packaging	the, of, ,, a, and, ., to, that, in, be	the, be, ., ,, and, in, of, a, to, have	

Some have questioned why the corpora at English-Corpora.org do not have the same wide range of association measures. The only one that we use is Mutual Information (MI) score, and we use it in a quite different way than others do. So are we doing something wrong?

We believe that our approach provides at least as good (if not better) results than sites with a supposedly more "sophisticated" approach, and which offer 5-10 different association measures. As evidence for this, take a look at the entries in the table above, for eight words -- three nouns from different frequency ranges, two adjectives, two verbs, and an adverb.

The first three rows show the top twelve collocates in iWeb[2]. Rows #1-2 are simply based on the frequency of the collocate, with a "Mutual Information" filter (of either 3.5 in row #1 or 6.0 in row #2). Row #3 sorts by Mutual Information score.[3]

Compare the collocates in these two rows to the collocates in the next seven rows from Sketch Engine (#4-10: LogDice to T-score)[4]. We would argue that our approach produces at least as good of results as those from Sketch Engine, probably better.

Most people can see right away that (Sketch Engine) [MI-score; row #7] gives strange, low frequency collocates. The top collocates from [MI3; row #8] are better than with [MI], but soon they peter out, and then they give way to high frequency words (*the, to, with*, etc). Both [Log Likelihood] and [T-score] (rows #9-10) produce very poor lists of collocates -- mainly just high frequency words, which provide little insight into the meaning and usage of the node word.

The other three statistics -- [LogDice], [MI.log-f], and [Minimum sensitivity] (#4-6) are quite good. But notice that our approach (which uses just raw frequency and

[2] iWeb is the corpus at English-Corpora.org that is most comparable to the corpora from Sketch Engine that we will be describing below, and which are based on tens of millions of web pages.

[3] To generate collocates in iWeb and COCA, select WORD from the search form, enter the word, and then click on Collocates on the following page. For all other corpora from English-Corpora.org, just click on Collocates in the search form on the first page and then enter the node word.

[4] (Details on the Sketch Engine search) The data is taken from a 38 billion word corpus of web pages. To obtain these collocates, we searched for concordance lines for the node word (as a lemma), with the indicated part of speech (e.g. the lemma *telescope* as a noun). We then clicked on the [Collocations] icon in the upper right-hand corner. To produce the collocates, we selected Attribute = [lempos] lower case; Span 4 Left to 4 right; Min Freq in Corpus = 50; Min freq in range = 20).

MI filter (rows #1-2; the results in yellow) often provides collocates that refer to basic concepts related to the node word, but which are missing (or much lower ranking) in #4-6.

For example, our collocates include {space, sky} for *telescope*, {painting, wall, chapel, church} for *fresco*, {tea, brew, fermented} for *kombucha*, {light, heat, visible} for *infrared*, and {material, plastic, reusable} for *recyclable.*

In addition, the collocates in #4-6 have many more proper nouns, which may not provide much insight into the meaning of the node word, unless you already know what they refer to {e.g. *cherenkov, spitzer, jacque, brancacci, kevita, scoby, health-ade*}. They also contain many lower frequency words that would probably not help a non-native speaker understand the meaning of the node word {*thermography, spectrometer, polypropylene, myogenic, otoacoustic, somatosensory,* etc).

So how do we obtain the quality of the collocates with the corpora at English-Corpora.org? To generate collocates in our approach, we do things almost backwards from how others do it.

1. First, we use frequency to find collocates -- raw frequency -- the most common nouns, verbs, adjectives, and adverbs that occur near the node word -- and even function words like *the, to, his*, etc.

2. Second, we use Mutual Information -- not to sort the results, but simply to serve as a sort of "filter" -- to *filter out* high frequency words. In row #1 we use an MI score of 3.5, and in row #2 we use an MI score of 6.0.

That's it. With those two steps, we produce collocates of the quality seen in #1-2 above.

There are at least three other important advantages of our “simple” but very effective approach:

1. Statistical tests #4-10 are available at Sketch Engine, but only barely. Even for low-frequency words like the eight words above, it takes anywhere between 4-8 *minutes* to generate the collocates for each node word (with 7-8 minutes being the norm). At this rate, a user could find the collocates (using the tests shown above) for only 7 or 8 words in one hour. With iWeb (or COCA, which has the same functionality), it takes less than one second to find the collocates -- in other words, our approach is 200-500 times as fast.[5]

2. For some tests like MI3, better results can be obtained by adjusting the values for settings like [Min Freq in Corpus] or [Min freq in range] to higher values in Sketch Engine. But this is often a function of corpus size (higher values for larger corpora). That assumes a lot of trial and effort on the part of users.

3. With the corpora from English-Corpora.org, users can quickly and easily adjust the [Mutual Information] setting to generate higher frequency / less specific collocates, or lower frequency / more specific collocates (as in rows #1-2 above). For example, the following are the top collocates for *telescope*, using different MI values:

MI	Comments	Collocates (examples)
1.0	more basic words	space, use, large, radio, small, image, look
3.0	a good starting point	space, large, radio, sky, optical, binoculars, observatory
6.0	more specific to the node word	optical, binoculars, observatory, astronomer, infrared, eyepiece, ground-based, aperture
8.5	very specific; similar to tests #4-6 above	ground-based, kepler, refractor, hubble, keck, celestron, space-based, gamma-ray, focuser, earth-based, collimate
For those who want even more specific collocates, just use simple Mutual Information, as in our row #3 (with a frequency threshold).		

Maybe you are still not convinced, and you think that we really do need all of the association measures #4-10 in the table above. Here is a challenge: If someone could provide us with 9-10 consecutive words from this frequency list where any of the tests in rows #4-6 provide better collocates than our approach (#1-2 above, in yellow), then we would incorporate these additional statistical tests into English-Corpora.org. But after looking at the collocates for a wide range of node words at Sketch Engine and English-Corpora.org during the last 3-4 years, we don't believe that will be necessary.

Our approach is more simple, and it provides better results.

[5] Sketch Engine also has a [Word Sketch] function from the main page and it is very fast (like our approach). But we are referring to the tests #4-10 above.