Constructions and long-tailed distribution of their collexemes: A Look at the relationship between low-frequency words and constructional prototypes

Yoichiro Hasebe (Doshisha University)

What is "Long-tail"?

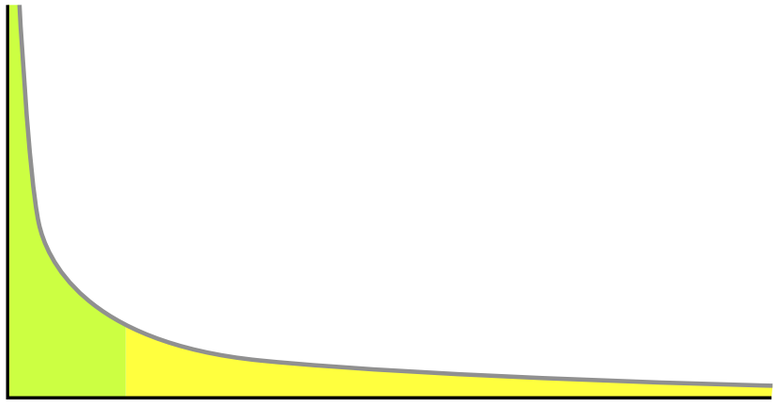
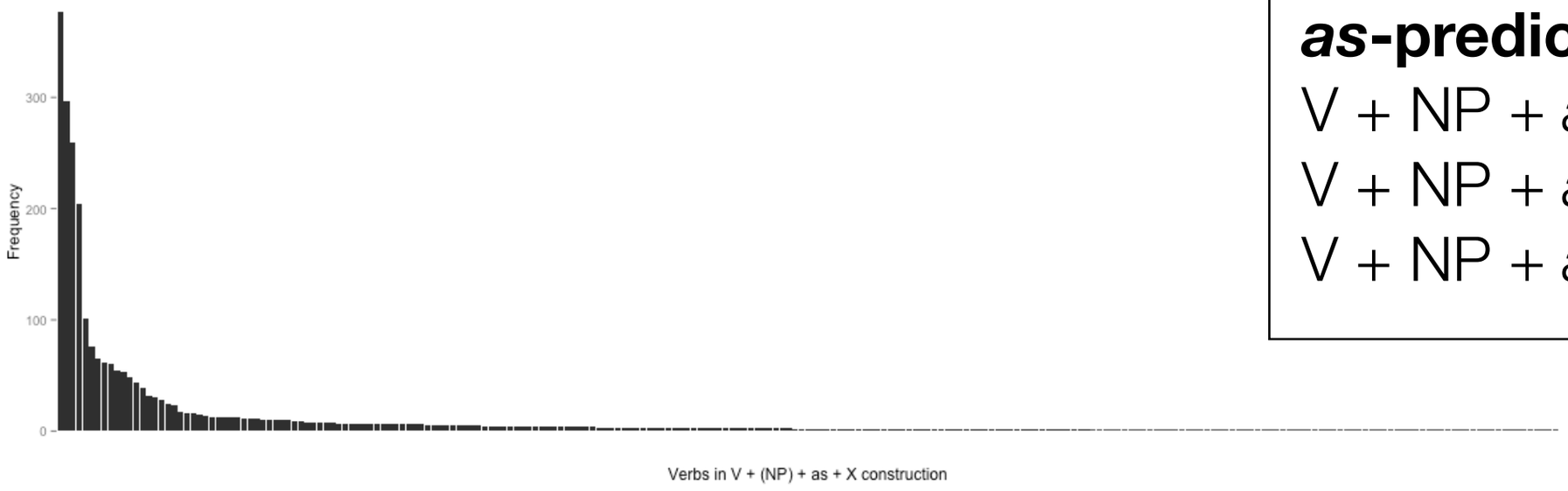


Figure 1: Long tail distribution



***as*-predicative**
V + NP + *as* + NP
V + NP + *as* + AdjP
V + NP + *as* + *-ing*

Figure 2: 2,632 instances of 239 types of collexemes of the *as*-predicative

What is great about collostructional analysis?

Table 1: Collexemes of the *as*-predicative (ICE-GB)

Rank	Verb	Word	Obs.	Col. Str.
1	***regard***	99	80	166.476
2	***describe***	259	88	134.87
3	***see***	1988	111	78.79
4	***know***	2120	79	42.796
5	***treat***	92	21	28.224
6	***define***	83	18	23.843
7	***use***	1228	42	21.425
8	***view***	41	12	17.861
9	***map***	23	8	12.796
10	***recognize***	114	12	12.159
11	***categorize***	10	6	11.525
12	***perceive***	28	6	8.304
13	***hail***	4	3	6.316
14	***appoint***	35	5	6.073
15	***interpret***	35	5	6.073
16	***class***	5	3	5.920
17	***denounce***	7	3	5.379
18	***dismiss***	25	4	5.158
19	***consider***	264	9	5.079
20	***accept***	178	7	4.467
21	***name***	41	4	4.282
22	***portray***	19	3	3.956
23	***advert to***	4	2	3.835
24	***diagnose***	6	2	3.440
25	***think of***	206	6	3.209
26	***depict***	8	2	3.172
27	***cite***	9	2	3.064
28	***rate***	9	2	3.064
29	***train***	40	3	2.981
30	***cast***	41	3	2.950

(Gries et al. 2005)

What is great about collostructional analysis?

> The attested usage of *regard* ... implies that, once this verbal item is known, it will most strongly be associated with the as-predicative, representing, as it were, the compressed version of the construction's semantics.
>
> (Gries et al. 2005: 652)

Construction as a complex network

Points that I would like to make regarding collostructional analysis:

- **The sub-types of the construction** and the relationship between the sub-types must be appropriately considered.
- It is not necessarily the case that there is only one prototypical collexeme; there could be **multiple prototypes** that are complementary to each other.

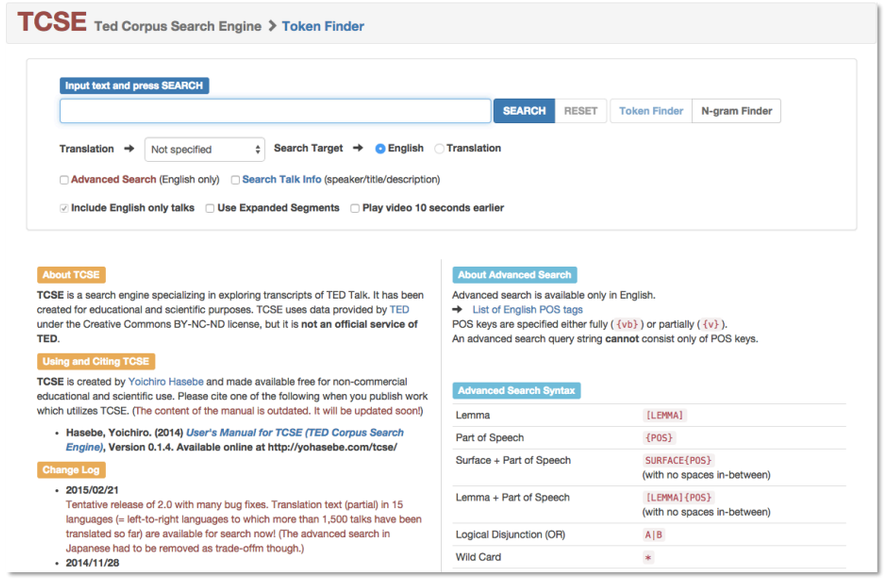
A construction may be characterized by a long-tailed graph of its collexemes, but behind such a simplistic facade, there is a complex network of sub-types of the construction.

TED as a corpus

TED Corpus Search Engine (TCSE)

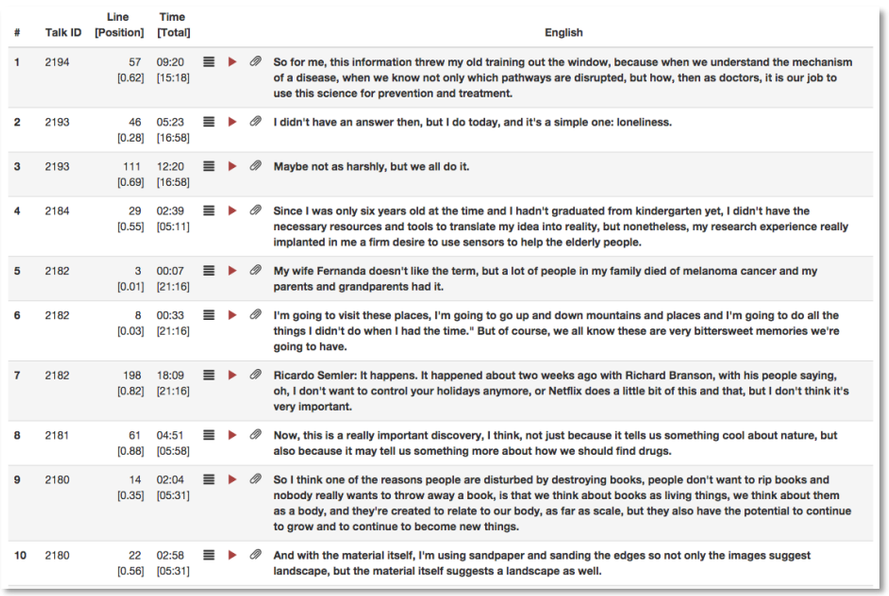
http://yohasebe.com/tcse

- Contains about 4,000,000 words from 1,900 talks of TED
- All text is POS tagged and search keys can include surface forms, lemmas, POS tags, etc.



ICE-GB (1 million)
687 instances of the *as*-predicative
with 107 verb types

TED Corpus (4 million)
2,632 instances of the *as*-predicative
with 239 verb types



#	Talk ID	Line [Position]	Time [Total]	English
1	2194	57 [0.62]	09:20 [15:18]	So for me, this information threw my old training out the window, because when we understand the mechanism of a disease, when we know not only which pathways are disrupted, but how, then as doctors, it is our job to use this science for prevention and treatment.
2	2193	46 [0.28]	05:23 [16:58]	I didn't have an answer then, but I do today, and it's a simple one: loneliness.
3	2193	111 [0.69]	12:20 [16:58]	Maybe not as harshly, but we all do it.
4	2184	29 [0.55]	02:39 [05:11]	Since I was only six years old at the time and I hadn't graduated from kindergarten yet, I didn't have the necessary resources and tools to translate my idea into reality, but nonetheless, my research experience really implanted in me a firm desire to use sensors to help the elderly people.
5	2182	3 [0.01]	00:07 [21:16]	My wife Fernanda doesn't like the term, but a lot of people in my family died of melanoma cancer and my parents and grandparents had it.
6	2182	8 [0.03]	00:33 [21:16]	I'm going to visit these places, I'm going to go up and down mountains and places and I'm going to do all the things I didn't do when I had the time." But of course, we all know these are very bittersweet memories we're going to have.
7	2182	198 [0.82]	18:09 [21:16]	Ricardo Semler: It happens. It happened about two weeks ago with Richard Branson, with his people saying, oh, I don't want to control your holidays anymore, or Netflix does a little bit of this and that, but I don't think it's very important.
8	2181	61 [0.88]	04:51 [05:58]	Now, this is a really important discovery, I think, not just because it tells us something cool about nature, but also because it may tell us something more about how we should find drugs.
9	2180	14 [0.35]	02:04 [05:31]	So I think one of the reasons people are disturbed by destroying books, people don't want to rip books and nobody really wants to throw away a book, is that we think about books as living things, we think about them as a body, and they're created to relate to our body, as far as scale, but they also have the potential to continue to grow and to continue to become new things.
10	2180	22 [0.56]	02:58 [05:31]	And with the material itself, I'm using sandpaper and sanding the edges so not only the images suggest landscape, but the material itself suggests a landscape as well.

Collostructional analysis on the TED corpus

Table 2: Collexemes of the *as*-predicative (TED Corpus)

Rank	Verb	Word	Obs.	Col. Str.
1	***think of***	991	377	Inf
2	***refer to***	115	60	61.903
3	***describe***	504	101	56.010
4	***view***	140	49	40.129
5	***define***	315	55	27.613
6	***regard***	41	24	26.829
7	***perceive***	154	39	26.087
8	***treat***	506	62	22.439
9	***use***	6411	297	19.786
10	***look upon***	8	7	10.128
11	***dismiss***	32	10	8.171
12	***herald***	7	5	6.459
13	***brand***	13	6	6.282
14	***identify***	276	23	5.728
15	***recognize***	492	31	4.903

Rank	Verb	Word	Obs.	Col. Str.
16	***cite***	23	6	4.612
17	***list***	62	9	4.418
18	***write off***	4	3	4.129
19	***redefine***	59	8	3.776
20	***classify***	32	6	3.748
21	***conceive***	46	7	3.689
22	***construe***	6	3	3.448
23	***reframe***	15	4	3.264
24	***talk of***	9	3	2.851
25	***label***	63	7	2.845
26	***envision***	34	5	2.707
27	***dub***	10	3	2.704
28	***characterize***	40	5	2.390
29	***certify***	26	4	2.326
30	***manifest***	27	4	2.265

Categories of collexemes of the *as*-predicative

(1) a. **Mental verbs**
regard, know, recognize, consider, think of

b. **Speech-act verbs**
describe, define, portray, hail, denounce, depict

c. **Classification verbs**
categorize, class, diagnose

d. **Verbs of ascription of a role or status**
appoint, nominate, adopt, establish

e. **Verbs of provisional ascription of properties**
use, treat

(Gries et al. 2005: 653)

Speech act scenario and control cycle

Speech act scenario (Langacker 2008)

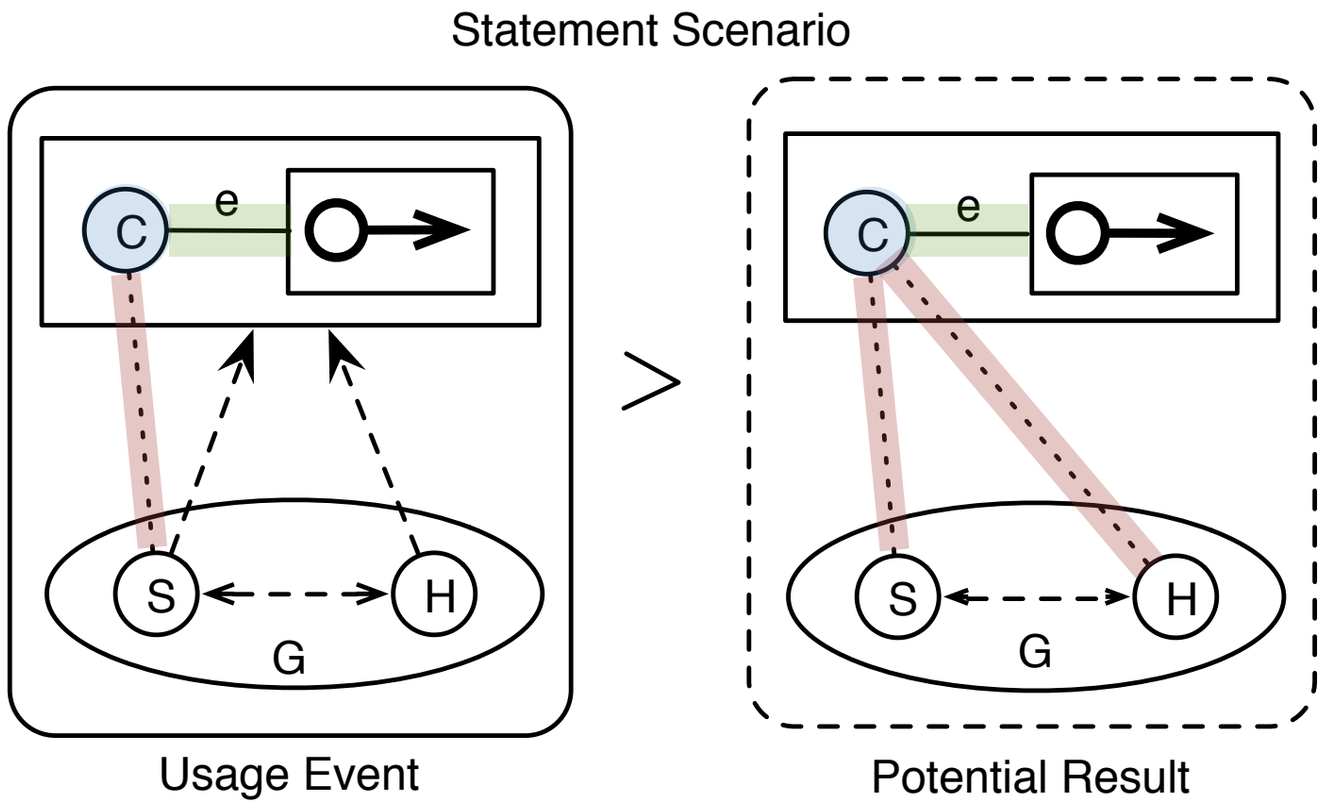


Figure 3: Speech-act scenario (Langacker 2008: 474)

Speech act scenario and control cycle

Control Cycle (Langacker 2009)

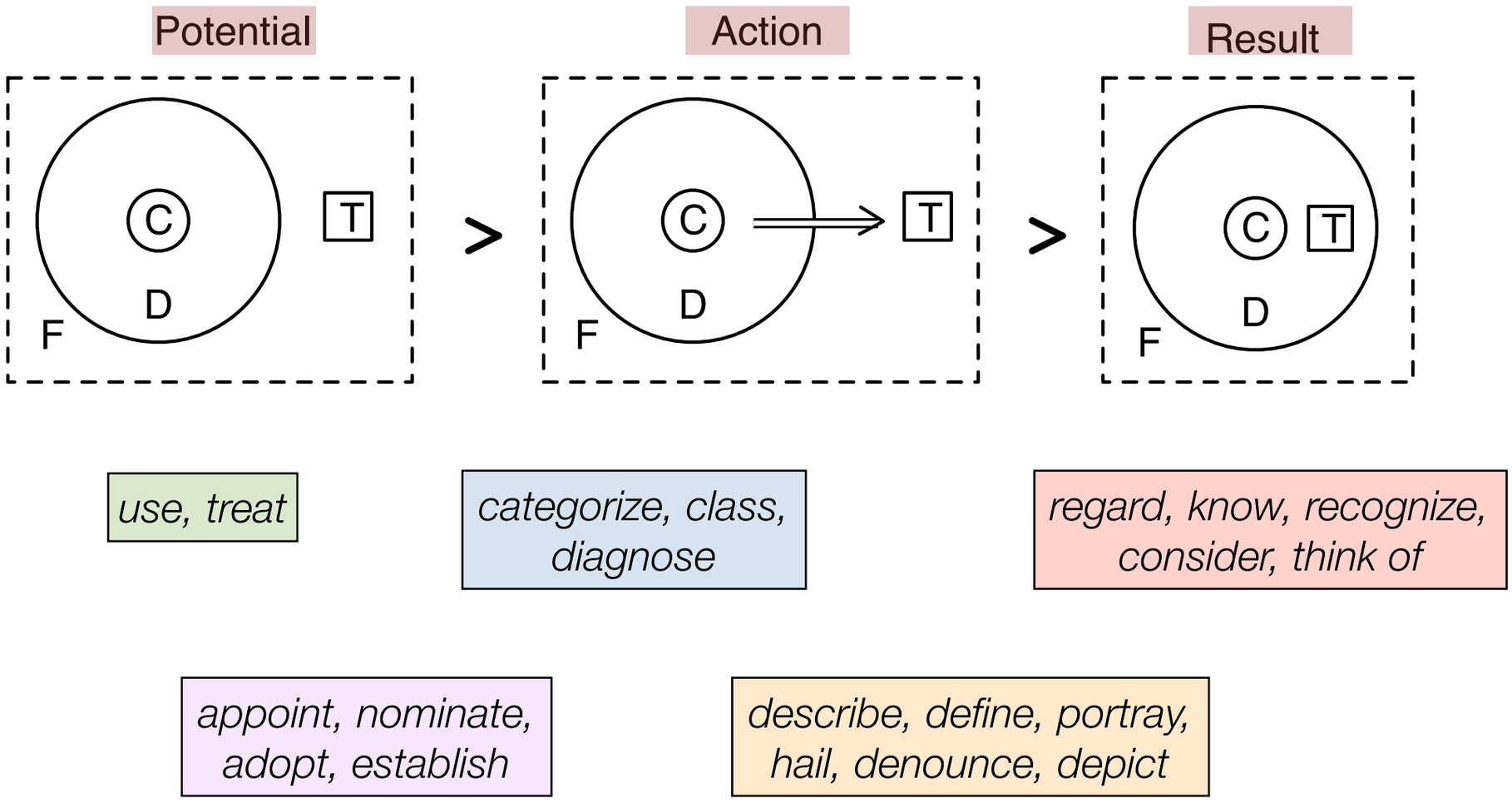


Figure 4: Control cycle and collexemes of the as-predicative

Groups of collexemes of the *as*-predicative

Table 2: Collexemes of the *as*-predicative (TED Corpus)

Rank	Verb	Word	Obs.	Col. Str.
1	***think of***	991	377	Inf
2	***refer to***	115	60	61.903
3	***describe***	504	101	56.010
4	***view***	140	49	40.129
5	***define***	315	55	27.613
6	***regard***	41	24	26.829
7	***perceive***	154	39	26.087
8	***treat***	506	62	22.439
9	***use***	6411	297	19.786
10	***look upon***	8	7	10.128
11	***dismiss***	32	10	8.171
12	***herald***	7	5	6.459
13	***brand***	13	6	6.282
14	***identify***	276	23	5.728
15	***recognize***	492	31	4.903

Rank	Verb	Word	Obs.	Col. Str.
16	***cite***	23	6	4.612
17	***list***	62	9	4.418
18	***write off***	4	3	4.129
19	***redefine***	59	8	3.776
20	***classify***	32	6	3.748
21	***conceive***	46	7	3.689
22	***construe***	6	3	3.448
23	***reframe***	15	4	3.264
24	***talk of***	9	3	2.851
25	***label***	63	7	2.845
26	***envision***	34	5	2.707
27	***dub***	10	3	2.704
28	***characterize***	40	5	2.390
29	***certify***	26	4	2.326
30	***manifest***	27	4	2.265

Words of low frequencies and constructional schema

Table 2': Collexemes that are less frequent and low in collostruction strength)

Rank	Verb	Word Freq.	Obs. Freq.	Relation	Col. Strength
195	**assume**	212	2	repultion	1.117
232	**confirm**	53	1	repultion	0.233
206	**launch**	267	4	repultion	0.800

Words of low frequencies and constructional schema

(2) a. So we don't have to **assume these principles as separate metaphysical postulates**.

b. And as you can see from the visuals, the service was responding and rescuing victims from the incident locations even before the police could cordon off the incident locations and formally **confirm it as a terror strike**.

c. We were also playing with SMS at the time at Odeo, so we kind of put two and two together, and in early 2006 we **launched Twitter as a side project at Odeo**.

Phases in control cycle and *as* + *-ing* construction

(3) a. Once, toward the end of my stay, a student said to me, "Professor, we never **think of you as being different from us**."

b. They were **all perceived as being less than normal** in all those characteristics -- more violent, etc. -- before the surgery.

c. A friend of mine **described it as standing in your own truth**, which I think is a lovely way to put it.

V + NP + *as* + *-ing*

- 372 instances with 76 different types of verbs
- No instance observed with "potential" verb *use*
- *Use* occurs with other types of the *as* predicative 297 times.

"Potential" phase and analogical expressions

(4) a. We know that cartoons can be **used as weapons**.

b. We're using the kind of skills that I've outlined: inner power -- the development of inner power -- through self-knowledge, recognizing and working with our fear, **using anger as a fuel**, cooperating with others, banding together with others, courage, and most importantly, commitment to active non-violence.

c. This allows you to pay it forward by **using this subject as a hook to science**, because SETI involves all kinds of science, obviously biology, obviously astronomy, but also geology, also chemistry, various scientific disciplines all can be presented in the guise of, "We're looking for E.T. "

Clustering verb types using Wordnet

Wordnet verb categories (Fellbaum 1998)

verb.body	verbs of grooming, dressing and bodily care
verb.change	verbs of size, temperature change, intensifying, etc.
verb.cognition	verbs of thinking, judging, analyzing, doubting
verb.communication	verbs of telling, asking, ordering, singing
verb.competition	verbs of fighting, athletic activities
verb.consumption	verbs of eating and drinking
verb.contact	verbs of touching, hitting, tying, digging
verb.creation	verbs of sewing, baking, painting, performing
verb.emotion	verbs of feeling
verb.motion	verbs of walking, flying, swimming
verb.perception	verbs of seeing, hearing, feeling
verb.possession	verbs of buying, selling, owning
verb.social	verbs of political and social activities and events
verb.stative	verbs of being, having, spatial relations
verb.weather	verbs of raining, snowing, thawing, thundering

Clustering verb types using Wordnet

Table 3: Wordnet categories and verb types in the *as*-predicative in TED Corpus

Wordnet Category	Num. V-types	Obs. Freq.	Verb Types
verb.cognition	42	926	*think of, view, regard, identify*
verb.perception	9	419	*perceive, look upon, display, present, show*
verb.communication	58	354	*refer to, describe, dismiss, herald*
verb.consumption	2	298	*use, utilize*
verb.social	21	175	*treat, brand, register, boycott*
verb.possession	15	132	*dispense, consign, adopt, sell*
verb.stative	18	113	*define, redefine, personify, uphold*
verb.creation	26	102	*recast, model, erect, reinvent*
verb.contact	20	49	*reframe, throw back, entrench, deposit*
verb.motion	7	31	*usher in, ship, bring, run*
verb.change	12	20	*accrue, devalue, condense*
verb.emotion	4	6	*revere, warship, want, like*
verb.competition	4	6	*battle, protect, attack, play with*
verb.body	1	1	*revive*
verb.weather	0	0	

Verb groups of "potential" phase

Table 3': Wordnet categories and the "potential" phase in the control cycle

Wordnet Category	Num. V-types	Obs. Freq.	Verb Types
verb.consumption	2	298	*use, utilize*
verb.creation	26	102	*recast, model, erect, reinvent*
verb.motion	7	31	*usher in, ship, bring, run*
verb.change	12	20	*accrue, devalue, condense*
verb.emotion	4	6	*revere, warship, want, like*
verb.competition	4	6	*battle, protect, attack, play with*
verb.body	1	1	*revive*

Verb groups of "potential" phase

(5) a. So we were **using the body as really the catalyst** to help us to make lots of new bone.

b. And then at the back of the greenhouse, it **condenses a lot of that humidity as freshwater** in a process that is effectively identical to the beetle.

c. Design something that makes oxygen, sequesters carbon, fixes nitrogen, distills water, **accrues solar energy as fuel**, makes complex sugars and food, creates microclimates, changes colors with the seasons and self-replicates.

d. I don't sleep that much, and I've come to this thing about, like, not sleeping much as being a great virtue, after years of kind of **battling it as being a terrible detriment**, or something.

Verb groups of "potential" phase

(5) a. Companies were created to limit financial risk, they were never intended to be **used as a moral shield**.

b. And so we saw an opportunity to bring design as this untouched tool, something that Bertie County didn't otherwise have, and to be sort of the -- to **usher that in as a new type of tool** in their tool kit.

c. We must **revive politics as the power** to imagine, reimagine, and redesign for a better world.

d. On Todagin Mountain, **revered by the Tahltan people as a wildlife sanctuary** in the sky, home to the largest population of stone sheep on the planet ...

Summary and conclusion

- The absolute ranks of collexemes should not be given too much significance:
  - Any corpus is not necessarily a perfect sample of the language
  - A construction may not be simplistic enough to allow one to define it in a linear scale.
- Instances of a construction are clustered into multiple groups
  - The groups may have natural connection as in the case of the *as*-predicative
  - Each group (or "phase") comprises a radial category having elements that are varied in their saliency and frequency.

To characterize such a complex nature of a construction as its entirety, it is imperative to look at those less frequent and less salient collexemes of the construction.

References

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gries, Stefan Th. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics. *Studies in Language* 11(3), 477-510.

Gries, Stefan Th., Beate Hampe, and Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4), 635-676.

Hasebe, Yoichiro. 2015. Design and implementation of an online corpus of presentation transcripts of TED Talks. Paper presented at the 7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond.

Langacker, Ronald. W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Langacker, Ronald W. 2009. *Investigations in Cognitive Grammar*. Berlin: Walter de Gruyter.

Schmid, Hans-Jörg and Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexico-grammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3), 531-577.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), 209-243.