

User's Manual for TCSE (TED Corpus Search Engine)

Version 0.1.7

Yoichiro Hasebe
Doshisha University

yohasebe@gmail.com

November 28, 2014

Contents

1	Introduction	2
1.1	What is TCSE?	2
1.2	About TED	2
1.3	Using and Citing TCSE	2
1.4	Acknowledgements	3
2	Token Finder	4
2.1	Basic Search	4
2.2	Showing Japanese Translation Text	7
2.3	Showing Expanded Segments	8
2.4	Advanced Search	8
2.4.1	Lemma	9
2.4.2	Parts of Speech	9
2.4.3	Other Advanced Search Options	10
2.5	Searching Talk Information	11
3	N-gram Finder	12
3.1	Basic Usage	12
3.2	Two Dispersion indices	14
4	Data Statistics	16
5	Frequently Asked Questions	17

Chapter 1

Introduction

1.1 What is TCSE?

TCSE is a search engine created by Yoichiro Hasebe (yohasebe@gmail.com) that specializes in exploring transcripts of TED Talk for educational and scientific purposes.¹ The working web application of this system is available at <http://yohasebe.com/tcse/>.

TCSE has been developed as an assistance tool for language learners/educators and linguistic researchers. Users can do the following:

- Search for segments of talk that match specified text strings in more than 1700 TED Talks;
- Study the context of talk segments in text, audio, and video formats;
- Input keywords (such as author, title, description) for easy retrieval of particular TED Talks;
- Discover frequent and/or characteristic phrasal expressions in TED.

1.2 About TED

The following description of TED (Technology, Education, and Design) is obtained from its official website:²

TED is a platform for ideas worth spreading. Started in 1984 as a conference where technology, entertainment and design converged, TED today shares ideas from a broad spectrum—from science to business to global issues—in more than 100 languages. Meanwhile, independent TEDx events help share ideas in communities around the world.

Contents of TED Talks are available under the Creative Commons BY-NC-ND license, which allows non-commercial entities.³ For further details, see the TED Talks Usage Policy.⁴

1.3 Using and Citing TCSE

Created by Yoichiro Hasebe, TCSE has been made freely available for non-commercial educational and scientific use. Please cite one of the following when using TCSE in your published work.

¹TCSE uses data provided by TED but is not an official service of TED.

²<http://www.ted.com/about/our-organization>

³<http://creativecommons.org/licenses/by-nc-nd/3.0/>

⁴<http://www.ted.com/about/our-organization/our-policies-terms/ted-talks-usage-policy>

- Hasebe, Yoichiro. (2014) *User's Manual for TCSE (TED Corpus Search Engine)*, Version 0.1.7. Available online at <http://yohasebe.com/tcse/> .
- Hasebe, Yoichiro (2014) 'Possibility of linguistics research of text in context using TED corpus.' Paper presented at the 18th Meeting of Tokyo Linguistic Colloquium.
- 長谷部陽一郎. (2014) 「TED コーパスを用いた文脈重視の言語分析の可能性」東京言語学コロキウム第 18 回研究会発表資料.

To report a bug in TCSE, use the following contact information:

- Yoichiro Hasebe (Doshisha University): yohasebe@gmail.com

Lastly, please do not forget to explicitly reference TED as the original source of the materials.

- TED: <http://ted.com>

1.4 Acknowledgements

I express sincere thanks to all the people involved in the TED and TEDx conferences for sharing their great insights as well as providing precious linguistic resources for education and research.

I greatly thank the following people who have supported and encouraged the development of TCSE:

- Jae-Ho Lee (University of Tsukuba)
- Haruo Nishinoh (Doshisha University)

Special thanks are also extended to the students who attended the author's seminar for test-running earlier versions of TCSE, held at the faculty of Global Communications, Doshisha University.

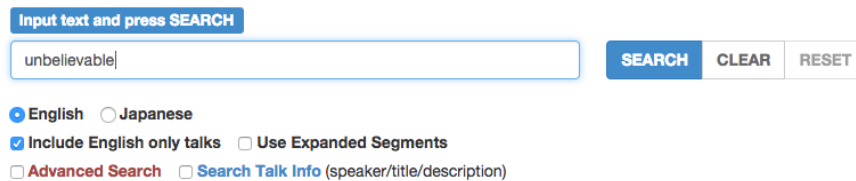
Chapter 2

Token Finder

The Token Finder function of TCSE searches for talk segments containing the text string specified in the search box and offers many options for accessing contextual data.

2.1 Basic Search

By default, TCSE conducts a basic search with `Advanced Search` unchecked. Type a search string into the text box, and press the `SEARCH` button (see Figure 2.1).



Input text and press SEARCH

unbelievable|

SEARCH CLEAR RESET

☒ English ☐ Japanese

☒ Include English only talks ☐ Use Expanded Segments

☐ Advanced Search ☐ Search Talk Info (speaker/title/description)

Figure 2.1: Search box and options

Figure 2.2 is an example of a search result. The text segments matching the input string are shown in descending order of `Talk ID`, which is assigned to each TED Talk. The larger the Talk ID, the newer the talk. If you hover the mouse cursor over the Talk ID, the title and the speaker of the talk will appear in a popup box.

Also, the line numbers and their relative positions (from 0 to 1) in the talk are shown in Figure 2.2. A TED Talk transcript consists of segments corresponding to lines of subtitles, which are originally intended to be sequentially shown on the video screen. Next to each line number is time of the segment (e.g., 00:50) and the total duration of the talk (e.g., 40:15). If the mouse is clicked on any of these items, a sub-window pops up showing the full text of the talk, with the queried segment highlighted, as shown in Figure 2.2.

Total number of items: 77

[Prev 100](#)
[Next 100](#)

#	Talk ID	Line [Position]	Time [Total]		English	Japanese
1	2055	18 [0.04]	00:50 [40:15]	☰ ▶ 🔗	unbelievable , right?	信じられないでしょう？
2	1990	277 [0.8]	11:15 [27:47]	☰ ▶ 🔗	She was an unbelievable role model.	素晴らしく模範的でした
3	1989	110 [0.26]	04:20 [36:33]	☰ ▶ 🔗	it was, like, unbelievable ,	信じられなかったけど
4	1964	501 [0.76]	18:50 [49:17]	☰ ▶ 🔗	that we are in an unbelievable situation	米国に住み 素晴らしい教育を
				☰ ▶ 🔗	and it has been unbelievable .	そこからは驚きの連続でした
				☰ ▶ 🔗	And we face unbelievable public safety challenges	これは治安上 極めて深刻な状況です
7	1890	57 [0.17]	02:44 [29:16]	☰ ▶ 🔗	one man with an unbelievable sense of his humanity.	驚くべき人間性を持った たった 1 人の人間による偉業です

Boyd Varty:
What I learned from Nelson Mandela

ボイド・ヴァーティ: ネルソン・マンデラが教えてくれたこと

Figure 2.2: Basic search results (excerpt)

Clicking on any of the three icons (set of horizontal lines, red triangle, and mini-clip) in each line of Figure 2.2 will bring up a sub-window containing contextual data (see Figure 2.3). The horizontal lines icon shows a sub-window containing paragraph text (Figure 2.4). The red triangle icon, as might expect, brings up a TED Talk video. The video automatically starts playing at the time location where the queried line is uttered (Figure 2.5). The mini-clip icon enables the user to copy the url of the video screen.

Nick Hanauer: Beware, fellow plutocrats, the pitchforks are coming			
Nick Hanauer is a rich guy, an unrepentant capitalist — and he has something to say to his fellow plutocrats: Wake up! Growing inequality is about to push our societies into conditions resembling pre-revolutionary France. Hear his argument about why a dramatic increase in minimum wage could grow the middle class, deliver economic prosperity ... and prevent a revolution.			
ニック・ハノーアー: 超富豪の仲間たち、ご注意を — 民衆に襲われる日がやってくる			
ニック・ハノーアーは金持ちであり、頑固な資本家でもあります。そして彼には超富豪の仲間たちに言いたいことがあります。「目を覚ませ！」格差の拡大により私たちの社会は革命前のフランスを髣髴とさせる状態に追いやられてしまいそうです。なぜ最低賃金の大幅な引き上げが中産階級の成長や経済的繁栄をもたらし、革命を阻止することができるのか。彼の主張を聞いてみましょう。			
		industries.	てきました
13	00:35	I was the first non-family investor in Amazon.com.	アマゾン社に非同族で出資したのは私が初めてでした
14	00:38	I cofounded a company called aQuantive	私はアクアンティブという会社を 共同で立ち上げ
15	00:41	that we sold to Microsoft for 6.4 billion dollars.	マイクロソフト社に64億ドルで 売却しました
16	00:45	My friends and I, we own a bank.	友人と共に銀行を所有しています
17	00:48	I tell you this — (Laughter) —	これをお伝えしたのは — (笑)
18	00:50	unbelievable, right?	信じられないでしょう?
19	00:52	I tell you this to show	これをお伝えしたのは私の人生が 他の多くの超富豪と
20	00:55	that my life is like most plutocrats.	同じだと言いたかったからです

Figure 2.3: Sub-window showing full text

Nick Hanauer: Beware, fellow plutocrats, the pitchforks are coming			
ニック・ハノーアー: 超富豪の仲間たち、ご注意を — 民衆に襲われる日がやってくる			
1	00:01	◇ You probably don't know me, ◇ but I am one of those .01 percenters ◇ that you hear about and read about, ◇ and I am by any reasonable definition a plutocrat. ◇ And tonight, what I would like to do is speak directly ◇ to other plutocrats, to my people, ◇ because it feels like it's time for us all ◇ to have a chat. ◇ Like most plutocrats, I too am a proud ◇ and unapologetic capitalist. ◇ I have founded, cofounded or funded ◇ over 30 companies across a range of industries. ◇ I was the first non-family investor in Amazon.com. ◇ I cofounded a company called aQuantive ◇ that we sold to Microsoft for 6.4 billion dollars. ◇ My friends and I, we own a bank. ◇ I tell you this — (Laughter) — ◇ unbelievable, right?	
		◇ 皆さんは私をご存じないでしょうが ◇ 私は皆さんがあちこちで耳にする ◇ 上位 0.01%の富裕層の一人で ◇ つまり紛れもないブルートクラット（超富豪 政治権力者）です ◇ 今日は私の仲間である 超富豪の人たちに向けて ◇ お話したいと思っています ◇ 私たち超富豪が話し合うべき時が ◇ 来たように思うからです ◇ 多くの超富豪と同様 私も 資本家であることを ◇ 誇りに思い 悪びれてもいません ◇ 私は様々な業界で30を超える会社を ◇ 個人や共同で設立したり 資金提供したりしてきました ◇ アマゾン社に非同族で出資したのは 私が初めてでした ◇ 私はアクアンティブという会社を 共同で立ち上げ ◇ マイクロソフト社に64億ドルで 売却しました ◇ 友人と共に銀行を所有しています ◇ これをお伝えしたのは — (笑) ◇ 信じられないでしょう?	

Figure 2.4: Sub-window showing paragraph text



	aQuantive	ち上げ
15	00:41 that we sold to Microsoft for 6.4 billion dollars.	マイクロソフト社に64億ドルで売却しました
16	00:45 My friends and I, we own a bank.	友人と共に銀行を所有しています
17	00:48 I tell you this — (Laughter) —	これをお伝えしたのは — (笑)
18	00:50 unbelievable, right?	信じられないでしょう？
19	00:52 I tell you this to show	これをお伝えしたのは私の人生が他の多くの超富豪と
20	00:55 that my life is like most plutocrats.	同じだと言いたかったからです

Figure 2.5: Sub-window showing TED Talk video

2.2 Showing Japanese Translation Text

Where available, TCSE shows both the original English transcript and the Japanese translated version. Japanese transcripts can also be searched for tokens of a specified expression. To use this feature, check `Japanese` in the options. Some of the TED Talks are not yet translated into Japanese (see Figure 2.6). To exclude segments with no Japanese translation from the TCSE search results, uncheck `Include English only talks` before the search.

11	1779	303 [0.81]	12:49 [31:54]	≡ ▶ 🔗	Unbelievable improvements in efficiency.	
12	1753	397 [0.92]	16:05 [34:42]	≡ ▶ 🔗	It's unbelievable , really. But when you get into it,	足を踏み入れれば分かります
13	1642	393 [0.96]	18:46 [39:05]	≡ ▶ 🔗	costs an unbelievable amount of money.	
14	1614	75 [0.46]	03:43 [17:47]	≡ ▶ 🔗	My God, it's unbelievable .	本当にとても信じられない経験でした
15	1564	158 [0.77]	07:24 [18:51]	≡ ▶ 🔗	unbelievable .	議論がはじまったのです

Figure 2.6: Availability of Japanese translation

2.3 Showing Expanded Segments

As demonstrated above, TCSE search results are displayed as talk segments, which correspond to subtitle lines in the video. However, a more useful text unit is sometimes required. For instance, linguistic researchers often analyze text on a sentence-by-sentence basis. However, since many of the sentences in TCSE are fragmented into separate segments. Therefore, TCSE offers the `Use Expanded Segments` option. With this option enabled, the search results combine the segments so that no sentence is cut-off midway. Note that an expanded segment may not correspond to a single sentence. A TED Talk segment sometimes contains a boundary between two sentences (i.e., a full stop separating two sentences). In this case, the expanded segment comprises two or more sentences. A search result with the `Use Expanded Segments` option checked is shown in Figure 2.7.

3	1989	37 [0.22]	04:14 [36:33]	  	I was so nervous and I was thinking just, like, oh my God, oh my God, and reminding me, because I've had, like, some very, especially since the last time we were here at TED, it was, like, unbelievable , and then right after that, like, so many crazy things happened, like, we ended up going to the White House to perform.	ずっと緊張してて どうしよう どうしようって 思ってたけど そういえば 前回 TEDに出られたのも 信じられなかったけど それ以来 スゴいことが 立て続けに起きて 結局 — ホワイトハウスにも行ったのよ
4	1964	183 [0.75]	18:45 [49:17]	  	As they get older, they so know that our family belief is about responsibility, that we are in an unbelievable situation just to live in the United States and have a great education, and we have a responsibility to give back to the world.	大きくなった子供たちには 我が家の信念が「責任」であると よくわかっていきます 米国に住み 素晴らしい教育を受けられるだけでも 信じられない境遇であり 世界に還元する責任があるのです
5	1964	200 [0.82]	20:40 [49:17]	  	MG: Totally stunned. BG: We had never expected it, and it has been unbelievable .	メリンダ：本当に驚いたわ ビル：想像だにしていなかったからね そこからは驚きの連続でした
6	1914	46 [0.42]	04:59 [24:38]	  	And we face unbelievable public safety challenges because we have a situation in which two thirds of the people in our jails are there waiting for trial.	これは治安上 極めて深刻な状況です というのも拘留所に 収容されている人間の 3分の2は 裁判の開始を 待っているだけなのです
7	1890	17 [0.12]	02:39 [29:16]	  	He was bringing peace to a divided and violent South Africa, one man with an unbelievable sense of his humanity.	マンデラ氏は分断され暴力にまみれた南アフリカに平和をもたらしました 驚くべき人間性を持った たった1人の人間による偉業です
8	1883	12 [0.21]	00:51 [11:12]	  	The hate mail I get is unbelievable .	信じられないような 嫌がらせメールが来ます

Figure 2.7: Expanded segments

2.4 Advanced Search

Besides the basic search described above, TCSE offers an advanced search for text tokens. By checking the `Advanced Search` option, you can specify certain attributes expected in all items of the retrieved text. More specifically, in an advanced search query, you can request linguistic concepts, such as lemmas and parts of speech.

As in the basic search, a sequence of words (a phrase) can be specified in an advanced search. Not surprisingly, the single-space character is recognized as the sign separating words in a phrase. In other words, if you insert a space between two text strings, the strings are considered as separate words comprising the phrase.

2.4.1 Lemma

A lemma is the canonical form of a set of words. Thus *hunt*, *hunts*, *hunted*, and *hunting* are all variations of the same lemma *hunt*. In TCSE, a lemma is represented by brackets (e.g., [hunt]).

2.4.2 Parts of Speech

To retrieve the parts of speech (POS) information in each of the talk contents, the TED text in TCSE is parsed by Enju 2.4.2.¹ The resulting POS data are represented by two-letter (case insensitive) codes, as shown in Table 2.1. In an advanced TCSE search, you can specify a word with a certain POS using curly brackets (e.g., {vb} specifies a verb).

Table 2.1: Parts of speech used in TCSE

POS	Description
cc	Coordinating conjunction
cd	Cardinal number
dt	Determiner
ex	Existential there
fw	Foreign word
in	Preposition or subordinating conjunction
jj	Adjective
ls	List item marker
md	Modal
nn	Noun
pd	Predeterminer
po	Possessive ending
pr	Personal pronoun
rb	Adverb
rp	Particle
sy	Symbol
to	to
uh	Interjection
vb	Verb
wd	Wh-determiner
wp	Wh-pronoun
wr	Wh-adverb

A POS is specified by either its full two-letter code (e.g., {wd}, {wp}, and {wr}) or by the first letter of its code (e.g., {w}). The single code {w} includes all of its sub-types {wd}, {wp}, and {wr}. Thus, a search for {w} will simultaneously return Wh-determiners, Wh-pronouns, and Wh-adverbs, corresponding to {wd}, {wp}, {wr}, respectively.

Thus, a POS can be specified in a shortened form, such as {v} (verb), {n} (noun), {j} (adjective), and {r} (adverb), but the POS contents may be difficult to guess. In this case, it is useful to explore how TCSE analyzes sentences and identify the POS tags assigned to the words in question. To use this feature, click on an English transcript line in the Token Finder search results. A sub-window will appear with a table showing the lemma, the POS, the total frequency, and the frequency (per million words) of each word in the sentence (2.8). The same functionality is available for Japanese translation text (Figure 2.9).

¹<http://kmcs.nii.ac.jp/enju/?lang=en>

Surface	a	picture	is	worth	a	thousand	words	.
Lemma	a	picture	be	worth	a	thousand	word	-period-
POS	dt	nn	vb	jj	dt	cd	nn	.
Freq	89924	1055	56539	297	89924	327	912	209231
PerMil	20358.1674	238.8447	12800.0359	67.2387	20358.1674	74.0305	206.4704	47368.4413

Figure 2.8: Sentence statistics for a segment of English text

Surface	百聞	は	一見	に	しか	ず	です
Lemma	百聞	は	一見	に	しく	ぬ	です
POS	名詞-一般	助詞-係助詞	名詞-サ変接続	助詞-格助詞	動詞-自立	助動詞	助動詞
Freq	7	143715	30	116780	10	2400	75595
PerMil	1.8275	37520.5108	7.8323	30488.4337	2.6108	626.582	19736.0262

Figure 2.9: Sentence statistics for a segment of Japanese text

A POS representation can be combined with both a lemma representation and a surface representation (i.e., the realized form of a word). To do this, concatenate a POS representation with curly brackets to a lemma representation (with square brackets) or a surface representation (bare word form without brackets) without intervening space characters. For instance, `[help]{v}` corresponds to the verb *help* of any surface forms (i.e., *help*, *helps*, *helped*, and *helping*), and `helping{v}` corresponds strictly to the *helping* form of the verb *help*. In both cases, the noun representation of *help* is ignored.

2.4.3 Other Advanced Search Options

In an advanced search query, you can express logical disjunction (OR) by inserting a vertical bar between options (e.g., `apple|orange|banana`). You can also use a wildcard symbol (*) to retrieve two separated items. However, note that the wildcard retrieves text of any size within the (expanded) segment. Thus the search string `'my * idea'` will not only match *'my new idea'* and *'my crazy idea'*, but also *'my mum certainly wasn't very keen on the idea.'*

In linguistic research, it is sometimes necessary to specify the onset and the ending of a segment or expanded segment. In TCSE, a segment or expanded segment is opened by the special symbol ^ but is not closed by any special symbol, because a full stop (.), a question mark (?), or an exclamation mark (!) is sufficient. Table 2.2 shows some advanced search strings and examples of their possible matches.

Table 2.2: Examples of advanced search strings

Search String	Possible Matches
[excite]	<i>excite</i> <i>excites</i> <i>excited</i> <i>exciting</i>
{vb}	verb, any kind
to * surprise	<i>to our surprise</i> <i>to his surprise</i>
[read] {DT} [news paper article]	<i>they read these articles</i> <i>reading the paper or something</i> <i>I'm reading the news at six</i>
^ having {v}	<i>Having started the process</i> <i>Having said that</i>
[help]{n}	<i>an aunt offered financial help</i> <i>we called people for help</i>

2.5 Searching Talk Information

By checking the `Search Talk Info` option, you can search titles, speakers, and descriptions of the talks. Note that when this option is enabled, advanced search syntax is not available. Figure 2.10 shows the results of a talk information search for the keyword *environment*.










#	Talk ID	Speaker: Title	Description	Num of Elements	Num of Segments	Num of Expanded Segments	Duration
1	1926	   Leyla Acaroglu: Paper beats plastic? How to rethink <i>environmental</i> folklore	Most of us want to do the right thing when it comes to the <i>environment</i> . But things aren't as simple as opting for the paper bag, says sustainability strategist Leyla Acaroglu. A bold call for us to let go of tightly-held green myths and think bigger in order to create systems and products that ease strain on the planet.	166	445	166	35:30
		レイラ・アカログル: 紙はビニールに勝る? 環境の民間信仰を考え直す方法	多くの人が環境に良いことをしたいと思うでしょう。しかし実際は紙袋を選べば良いと言った単純なことではありません。				
2	945	   Johan Rockstrom: Let the <i>environment</i> guide our development	Human growth has strained the Earth's resources, but as Johan Rockstrom reminds us, our advances also give us the science to recognize this and change behavior. His research has found nine "planetary boundaries" that can guide us in protecting our planet's many overlapping ecosystems.	124	452	124	35:26
3	598	   Stewart Brand: 4 <i>environmental</i> 'heresies'	The man who helped usher in the <i>environmental</i> movement in the 1960s and '70s has been rethinking his positions on cities, nuclear power, genetic modification and geo-engineering. This talk at the US State Department is a foretaste of his major new book, sure to provoke widespread debate.	179	356	179	32:34
		スチュアート・ブランド「環境に関する4つの異端的考察」	1960年代、70年代の環境保護運動の先駆けとなった彼は、都市化、原子力発電、遺伝子組み換え食材、さらには地球工学に関しての彼自身の考え方について、再考を行いました。この米国国務省に向けてのスピーチで、世界で話題になるであろう彼の力作を垣間見る事が出来ます				

Figure 2.10: Results of a talk information search for *environment*

Chapter 3

N-gram Finder

3.1 Basic Usage

The N-gram Finder mode of TCSE offers a rather different view of the text in TED Talks. An n-gram is a sequence of linguistic units (i.e., morphemes, words, etc.) of n (2, 3, 4, ...) items. The n-gram concept is widely used in linguistics, information technology, and similar fields. By investigating the different frequencies of various n-grams, it should be possible to derive the linguistic sequences that are highly entrenched in the spoken language and those that are less common. To switch to N-gram Finder mode, click on the `N-gram` button (shown in Figure 3.1).



Figure 3.1: Switching to N-gram Search

Unlike Token Finder, the N-gram Finder mode accepts only the surface form of a single word at a time. For example, typing the word *read* and clicking the `SEARCH` button brings up clickable tab menus, as shown in Figure 3.2.

Dispersion ☒ Gries's DP ☐ Juilland's D

Word info

2-grams3-grams4-grams

	Surface	Lemma	POS	Freq	Num of Talks	Dispersion
1	read	read	{vb}	911	487	0.6681
2	read	read	{nn}	25	25	0.9834

Figure 3.2: Word Info in N-gram Finder mode

There are four tab menus in Figure 3.2: `Word Info`, `2-grams`, `3-grams`, and `4-grams`. The default selection is `Word Info`. The `Word Info` panel contains not the n-grams themselves, but the basic statistics of the surface form of the queried word. For example, given the query word *read*, the panel shows the lemma form of the word (*read*), and the part-of-speech ({vb} (verb) or {nn} (noun)). *Read* as a verb is used 911 times and is distributed among 487 talks;

as a noun, it is used 25 times and distributed among 25 talks. The numbers in the *Dispersion* column indicate how broadly and uniformly the word is distributed throughout the corpus. The default dispersion index is the Gries' *deviation of proportions* (DP). The smaller is the Gries' DP, the more is dispersed the word. For example, in Figure 3.2, *read* as a verb (DP = 0.6681) is more dispersed than *read* as a noun (DP = 0.9834).¹

Word info	2-grams	3-grams	4-grams
-----------	---------	---------	---------

→ 3-gram ALL	WORD	WORD	WORD
→ 3-gram #1	WORD	WORD	WORD
→ 3-gram #2	WORD	WORD	WORD
→ 3-gram #3	WORD	WORD	WORD

3-gram ALL						
	Word1	Word2	Word3	Freq	Num of Talks	Dispersion
1	{pr}	{md}	read	76	65	0.9527
2	{pr}	read	{pr}	47	41	0.9703
3	to	read	{pr}	39	34	0.976
4	you	can	read	32	30	0.9779
5	and	{pr}	read	31	28	0.9792
6	when	{pr}	read	28	28	0.9795
7	{pr}	read	the	26	24	0.9824
8	{md}	read	{pr}	26	24	0.983
9	to	read	the	25	23	0.9836
10	{pr}	read	about	21	19	0.9848

Figure 3.3: 3-grams containing *read*

By clicking on the tab menus, you can access data tables of 2-grams, 3-grams, and 4-grams. A 3-grams table of the word *read* is shown in Figure 3.3. The n-gram tables of TCSE collect the surface forms of text such as *you can read*. But very frequent word types such as {pr} (pronouns) and {md} (modals) are not only plainly n-grammed in TCSE but are also aggregated into single entries.

All of the items listed in n-gram tables are clickable. For instance, if you click on the fourth listed item in Figure 3.3, *you can read*, TCSE switches to Token Finder mode and returns the text tokens corresponding to the n-gram sequence (see Figure 3.4).

¹Gries' DP differs from many other dispersion indices, in that it decreases with increasing degree of dispersion (see next section).

#	Talk ID	Line [Position]	Time [Total]		English	Japanese
1	2061	23 [0.37]	01:44 [12:53]	☰ ▶ 🔍	But in fact, yeah, it's been a success, there's lots of things, Khan Academy for crying out loud, there's Wikipedia, there's a huge number of free e-books that you can read online, lots of wonderful things for education, things in many areas.	確かに成功していて いろいろなものがあります カーンアカデミーや ウィキペディアだってあります 無料で読めるオンライン書籍も 大量にあります 多くの教育に役立つ優れたものや 様々な分野のものがあります
2	1991	31 [0.43]	02:19 [11:53]	☰ ▶ 🔍	When someone hands an object to you, you can read intention in their eyes, their face, their body language.	誰かがあなたに何かを渡す時 その人の目や表情や ボディラングージから 意図を推し量れます
3	1820	118 [0.69]	11:45 [34:09]	☰ ▶ 🔍	By the end of the investigation, where you can read the full 27-page report at that link, we had photos of the cybercriminals, even the office Christmas party when they were out on an outing.	この調査を終える頃にはー このリンクから 27ページもある報告書を見られますがー サイバー犯罪者たちの写真も入手し 職場のクリスマスパーティーで 出かけたときの写真まであります 職場のクリスマスパーティーで 出かけたときの写真まであります
4	1587	73 [0.51]	08:09 [29:53]	☰ ▶ 🔍	So I'm not going to go through the whole details of the study because actually you can read about it, but the next step is observation. So here are some of the students doing the observations. They're recording the data of where the bees fly.	研究の詳細は 読むことができますので 細かくは話しませんが 次のステップは観察です 生徒が何人かで 観察をしています ミツバチがどこへ飛んだか 記録しています
5	1573	47 [0.46]	05:49 [25:58]	☰ ▶ 🔍	Now, the elements of the scene also communicate this to us, but you can read it straight off their faces, and if you compare their faces to normal faces, it would be a very subtle cue.	

Figure 3.4: Text Tokens containing the 3-gram *you can read*

3.2 Two Dispersion indices

The N-gram Finder of TCSE admits two dispersion indices, Gries' DP and Juilland's D. Although choice of dispersion index alters the order of n-grams, both indices show the extent to which the query word is dispersed among all talks in the corpus. A possible source of confusion is that Gries' DP assigns smaller numbers to more dispersed entities, whereas Juilland's D (like most dispersion indices) assigns larger numbers to more dispersed entities.

A brief overview of these dispersion indices may be helpful. Juilland's D was introduced by Juilland et. al. (1970) while searching for new ways to compile a frequency dictionary of the French language. Since then, many alternative dispersion indices have been proposed (for a summary, see Gries 2008), but Juilland's D remains among the most popular. In Juilland's D representation, the dispersion degree of a linguistic unit ranges from 0 (least dispersed throughout the corpus) to 1 (most dispersed). The formula, which is not presented here, yields fairly accurate results in many cases and is easily applied.

Gries (2008) pointed out several defects of Juilland's D, however. Most importantly, it does not account for the different text sizes of the files/sections in the corpus. Gries' DP dispersion index overcomes this defect and is thus utilized as the default in TCSE.

Although Gries' DP is superior to Juilland's D, both dispersion formulas produce similar results in most cases. When the results noticeably differ, it is understood that DP considers not only the total frequency of the n-grams and the number of files containing them, but also the different text sizes of files comprising the entire corpus. Figure 3.5 compares the top fifteen 3-grams containing the word *keep* ordered by DP (a) and D respectively (b).

	Word1	Word2	Word3	Freq	Num of Talks	Dispersion
1	to	keep	{pr}	104	94	0.8947
2	{pr}	{md}	keep	66	59	0.8648
3	to	keep	the	51	43	0.8388
4	{pr}	keep	{pr}	37	34	0.8243
5	and	{pr}	keep	33	32	0.8222
6	to	keep	it	27	26	0.8019
7	if	{pr}	keep	30	27	0.8013
8	going	to	keep	36	29	0.7934
9	keep	in	mind	23	23	0.7927
10	{md}	keep	{pr}	28	25	0.7869
11	keep	up	with	22	19	0.752
12	{md}	n't	keep	27	21	0.7471
13	to	keep	in	15	15	0.7428
14	to	keep	them	16	15	0.7358
15	want	to	keep	15	14	0.726

(a) ordered by DP

	Word1	Word2	Word3	Freq	Num of Talks	Dispersion
1	to	keep	{pr}	104	94	0.9345
2	{pr}	{md}	keep	66	59	0.9575
3	to	keep	the	51	43	0.9695
4	{pr}	keep	{pr}	37	34	0.975
5	and	{pr}	keep	33	32	0.9754
6	going	to	keep	36	29	0.9782
7	if	{pr}	keep	30	27	0.9808
8	to	keep	it	27	26	0.9815
9	{md}	keep	{pr}	28	25	0.9816
10	{md}	n't	keep	27	21	0.9847
11	keep	up	with	22	19	0.9849
12	keep	in	mind	23	23	0.987
13	to	keep	them	16	15	0.9882
14	keep	{pr}	from	18	15	0.9888
15	ca	n't	keep	21	15	0.9889

(b) ordered by DP

Figure 3.5: 3-grams containing the word *keep*

Chapter 4

Data Statistics

TCSE Version 0.1.7 contains the data of more than 1,700 TED Talks and will be regularly updated. The following data statistics are retrieved on November 28, 2014.

Table 4.1: Data statistics of English transcripts of TED Talks (v. 0.1.7)

Number of talks	1,769
Number of segments	498,374
Number of expanded segments	214,528
Number of elements	4,440,124
Number of lexical items	70,112

As mentioned in 2.4.2, TCSE analyzes the original English transcripts of TED Talks and assigns POS tags to them by parsing through Enju 2.4.2. The data in Table 4.1 are based on the POS tagging. Note that in TCSE, POS types are represented by two-letter codes, such as {vb} and {nn}; no distinction is made between subtypes, such as {vb} (verb, base form) and {vbd} (verb, past tense), or between {nn} (noun, singular or mass) and {nns} (noun, plural).

TCSE also uses the Japanese morphological analyzer MeCab 0.996 + IPA dictionary 2.7.0 to analyze Japanese translations of English transcripts. Thus, TCSE can conduct advanced searches of Japanese translation texts.¹ The data statistics of Japanese translation texts in TCSE are shown in Table 4.2.

Table 4.2: Data statistics of Japanese translation texts of TED Talks (v. 0.1.7)

Number of talks	1,548
Number of segments	422,260
Number of expanded segments	181,953
Number of elements	3,878,715
Number of lexical items	56,910

¹<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

Chapter 5

Frequently Asked Questions

Q1 Do you distribute TED Talk transcripts/translation text data?

A1 No, but project such as TED-LIUM (Rousseau et al., 2014)¹ and TED CLDC Corpus (Hermann and Blunsom, 2014)² provides downloadable packages containing transcript data from TED Talks.

Q2 Do you plan to update TCSE so that it includes newly released TED Talks? If so, how often?

A2 I plan to periodically update TCSE, but the schedule is not yet fixed. Please see the update information on the TCSE homepage (<http://yohasebe.com/tcse>).

Q3 Can TCSE handle case sensitive searches?

A3 No, TCSE searches are always case insensitive, in both regular search mode and advanced search modes.

Q4 Do you plan to extend TCSE to include translation data in languages other than Japanese?

A4 Not in the near future, although such an extension is certainly possible.

Q5 How can I submit a bug report? How can I suggest a new functionality to TCSE?

A5 Send an e-mail to Yoichiro Hasebe <yohasebe@gmail.com>.

¹<http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus>

²<http://www.clg.ox.ac.uk/tedcorpus>

Bibliography

- Gries, S. T. (2008). “Dispersions and adjusted frequencies in corpora”. *International Journal of Corpus Linguistics*, **13**(4) pp.403–437.
- Hasebe, Y. (2014). “Possibility of linguistics research of text in context using TED corpus”. Paper presented at the 18th Meeting of Tokyo Linguistic Colloquium, University of Tsukuba.
- Hermann, K. M. and P. Blunsom (2014). “Multilingual models for compositional distributional semantics”. *Proceedings of ACL*, <http://arxiv.org/abs/1404.4641>.
- Juilland, A. G., D. R. Brodin, and C. Davidovitch (1970). *Frequency dictionary of French words*. Berlin: Mouton de Gruyter.
- Rousseau, A., P. Deléglise, and Y. Estève (2014). “Enhancing the TED-LIUM Corpus with selected data for language modeling and more TED Talks”. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*.