CILC 2015 Valladolid

Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks

Yoichiro Hasebe

Doshisha University

Introduction

TED Corpus Search Engine (TCSE)

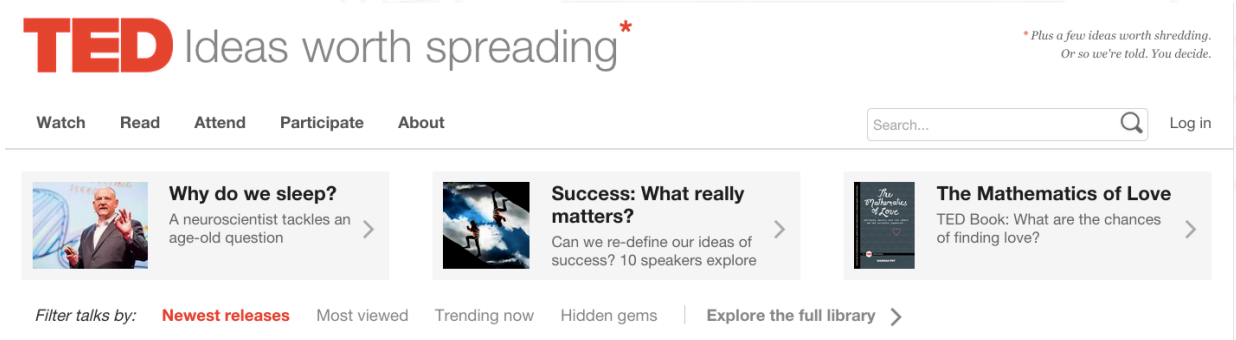
http://yohasebe.com/tcse

- Capable of searching TED Talks and show text/audio/video segments that match the input string
- Not only English transcripts but also their translations in 15 languages are available
- Designed and implemented on the usage-based model of language

About TED

TED (http://ted.com)

- Technology, Entertainment, and Design
- TED and TEDx conferences are held worldwide to share ideas worth spreading
- Speakers include artists, researchers, politicians, etc.
- Talk data are distributed under Creative Commons License（by-nc-nd 3.0）



Transcripts and translations

- More than 1,800 transcripts of TED talks are available online
- There are many volunteers transcribing, translating, and reviewing the talk text

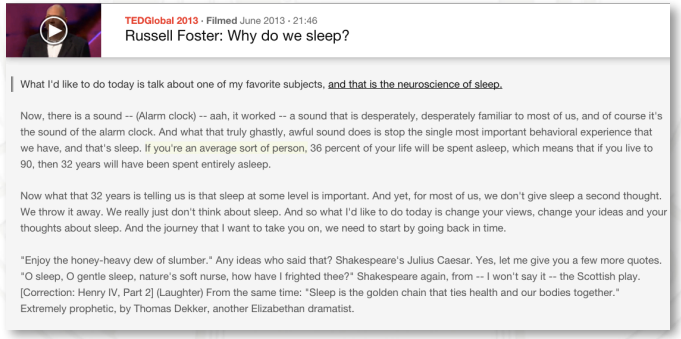
Translation stats

- 107 languages
- 19,807 translators
- 70,021 translations

(http://www.ted.com/participate/translate)

Limitations of official TED web system

- Text search functionality is not very sophisticated (no POS and lemma searches)
- Comparison between transcript and translation is not possible
- Comparison between translations is not possible
- Video/audio starts from the beginning (not from the segment that matches the input string)



TED as corpus

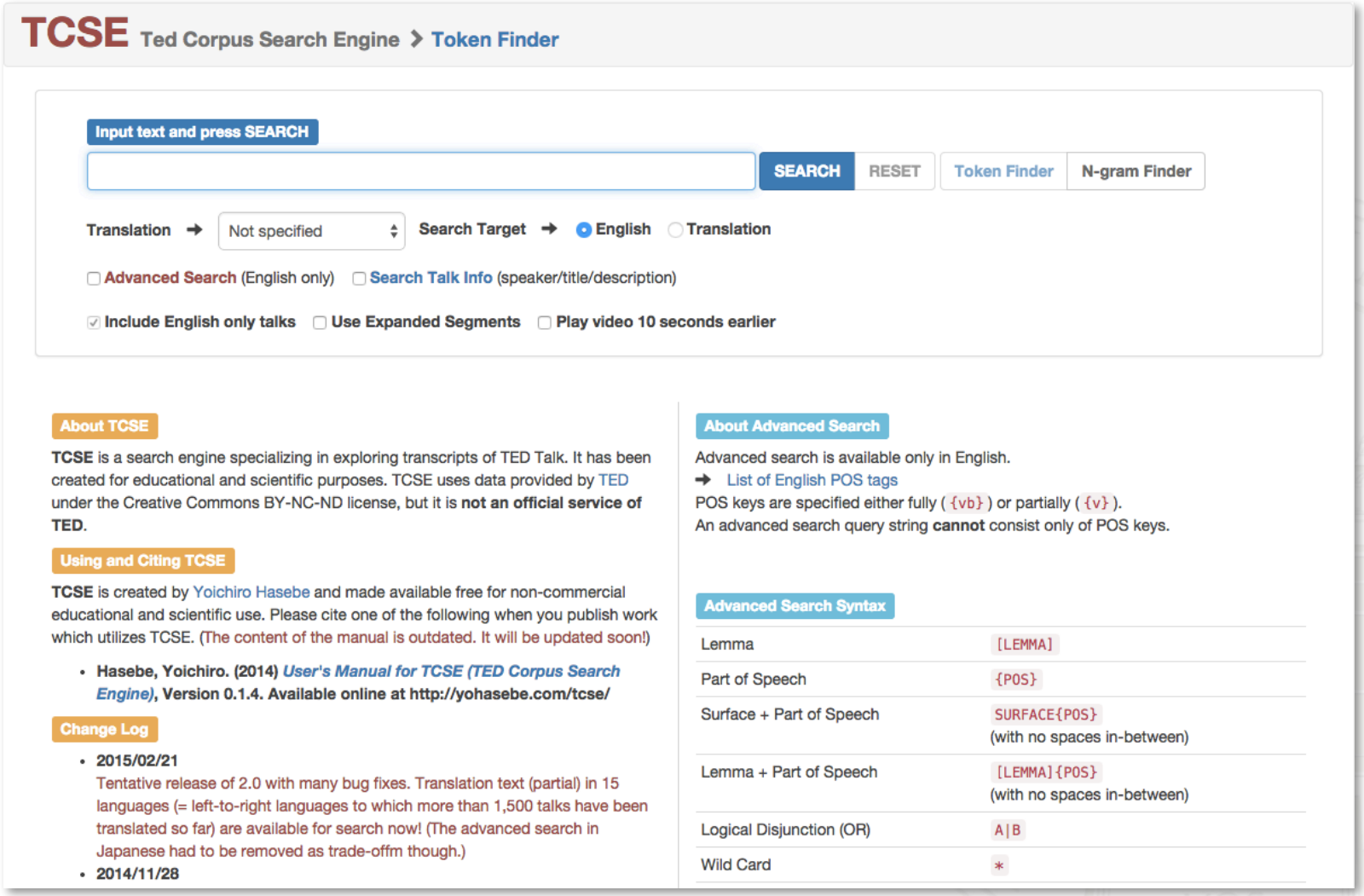
Cons

- It is not balanced.
- Token size is about 4,500,000 (not comparable to corpora such as BNC and COCA).
- It contains English by many non-native speakers.

Pros

- It is provides rich audio/video data.
- It can be considered as a corpus of presentation text.
- It contains wide varieties of English in the world.

http://yohasebe.com/tcse



Specifications of TCSE

Text units of different sizes

- segments
- expanded segments
- paragraphs

#	Talk ID	Line [Position]	Time [Total]	English
1	2194	149 [0.6]	09:28 [15:18]	when we know not only which pathways are disrupted, but how,
2	2193	90 [0.31]	05:23 [16:58]	I didn't have an answer then, but I do today,
3	2193	206 [0.72]	12:20 [16:58]	Maybe not as harshly, but we all do it.
4	2178	108 [0.33]	04:34 [14:39]	depends not on you, but on what you're trying to learn.
5	2178	209 [0.64]	09:07 [14:39]	is that listening to Mozart can make you not only cleverer but healthier, too.
6	2177	130 [0.41]	07:58 [18:50]	not receiving the support they needed, but increasingly isolated.
7	2176	8 [0.02]	00:20 [15:52]	This is not inevitable, but overcoming it requires diving deep
8	2176	175 [0.64]	10:15 [15:52]	not just the march, but the capacity signaled by that march, seriously.
9	2176	214 [0.78]	12:31 [15:52]	It's not the same game, but the differences are instructive.
10	2174	122 [0.5]	08:03 [15:34]	And that happens not only in Congo but also in many other conflict zones.

Specifications of TCSE

Text units of different sizes

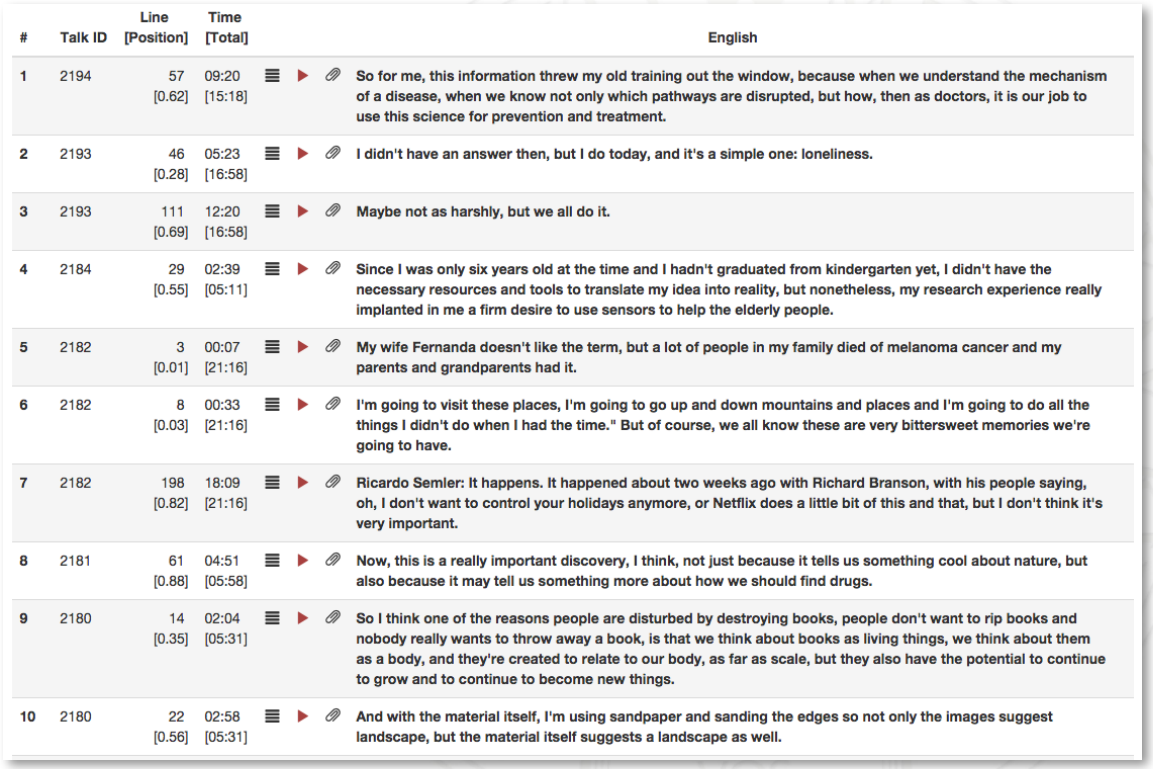
- segments
- expanded segments
- paragraphs

"not A but B" construction

[not] * but

975 segments

4189 expanded segments

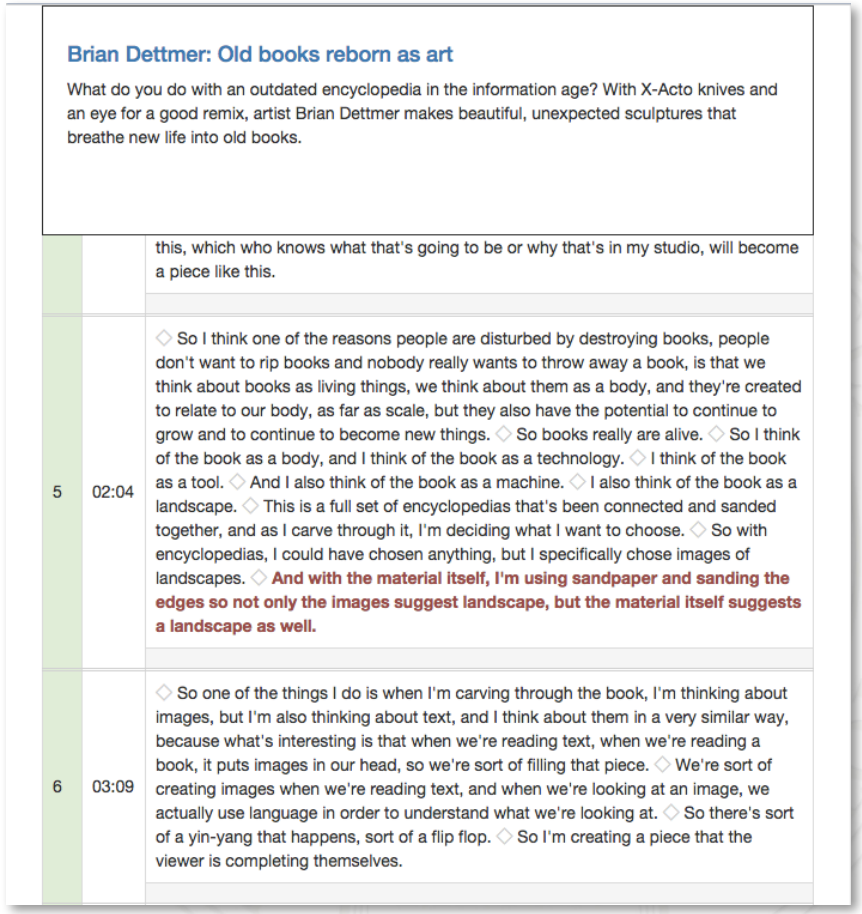


#	Talk ID	Line [Position]	Time [Total]	English
1	2194	57 [0.62]	09:20 [15:18]	So for me, this information threw my old training out the window, because when we understand the mechanism of a disease, when we know not only which pathways are disrupted, but how, then as doctors, it is our job to use this 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.

Specifications of TCSE

Text units of different sizes

- segments
- expanded segments
- paragraphs



Brian Dettmer: Old books reborn as art

What do you do with an outdated encyclopedia in the information age? With X-Acto knives and an eye for a good remix, artist Brian Dettmer makes beautiful, unexpected sculptures that breathe new life into old books.

		this, which who knows what that's going to be or why that's in my studio, will become a piece like this.
5	02:04	◇ 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. ◇ So books really are alive. ◇ So I think of the book as a body, and I think of the book as a technology. ◇ I think of the book as a tool. ◇ And I also think of the book as a machine. ◇ I also think of the book as a landscape. ◇ This is a full set of encyclopedias that's been connected and sanded together, and as I carve through it, I'm deciding what I want to choose. ◇ So with encyclopedias, I could have chosen anything, but I specifically chose images of landscapes. ◇ **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.**
6	03:09	◇ So one of the things I do is when I'm carving through the book, I'm thinking about images, but I'm also thinking about text, and I think about them in a very similar way, because what's interesting is that when we're reading text, when we're reading a book, it puts images in our head, so we're sort of filling that piece. ◇ We're sort of creating images when we're reading text, and when we're looking at an image, we actually use language in order to understand what we're looking at. ◇ So there's sort of a yin-yang that happens, sort of a flip flop. ◇ So I'm creating a piece that the viewer is completing themselves.

Specifications of TCSE

Regular search in English transcripts

1 (a) might as well
(b) as if
(c) far from

Regular search in translations (Japanese examples)

2 (a) かもしれない
(b) まるで
(c) 程遠い

Specifications of TCSE

Lemma search

3 (a) [make] sense
(b) [know] better
(c) [happen] to

Lemma and POS search

4 (a) [remember] to
(b) [remember] {v}
5 (a) [help]{n}
(b) [help]{v} (note: no space between [] and {})

Specifications of TCSE

Wild card and logical disjunction

6 [not] * but

7 [aslso] long as

Specification of segment onset position

8 (a) ^ again

(b) ^ still

(c) ^ now

Statistics of talks in TCSE

Number of talks	1,828
Total playing time of talks	396 hours
Total number of segments	511,923
Total number of expanded segments	220,565
Total number of word tokens	4,567,505
Total number of word types	80,790
Mean length of talks	13 minutes
Mean number of words	2,499
Mean words per minute	192

(as of February 21, 2015)

Translations

マチウ・リカール: 愛他性に導かれる生き方

愛他性とは何でしょうか？簡単に言えば、自分以外の人達の幸せを願うことです。そして、幸福学の研究者で仏教の僧侶でもあるマチウ・リカールが言うには、愛他性は短期的なものであれ長期的なものであれ、また仕事に関するものであれ暮らしに関するものであれ、何か決断を下す時に重要な指針として働くものなのです。

66	04:33	and the long term of the environment.	そして環境という 長期的なものです
67	04:37	When the environmentalists speak with economists,	環境活動家が経済学者と話をすると
68	04:39	it's like a schizophrenic dialogue, completely incoherent.	統合失調症風の会話になって 筋が全く通りません
69	04:42	**They don't speak the same language.**	**彼らは話す言語が 違っているんです**
70	04:45	Now, for the last 10 years, I went around the world	この10年 私は世界中を回って 至る所の
71	04:49	meeting economists, scientists, neuroscientists, environmentalists,	経済学者、科学者 神経科学者、環境活動家
72	04:53	philosophers, thinkers in the Himalayas, all over the place.	哲学者、ヒマラヤの思索家に 会ってきましたが
73	04:57	It seems to me, there's only one concept	先程の 3つの時間尺度の間を 取り持てるのは
74	05:01	that can reconcile those three time scales.	たった1つの概念しか ないように思えます
75	05:04	It is simply having more consideration for others.	それは単に 他者を もっと大事にするということです

ジョン・マクホーター 「テキスト・メッセージが言語を殺す （なんてね！）」

Surface	**and**	**look**	**at**	**what**	**language**	**really**	**is**	**,**
Lemma	and	look	at	what	language	really	be	-comma-
POS	cc	vb	in	wp	nn	rb	vb	,
Freq	132,819	4,387	16,236	23,713	915	8,763	52,336	247,679
PerMil	29,079	960	3,555	5,192	200	1,919	11,458	54,226

Transcript and Translation

English	and look at what language really is,
Chinese, Simplified	好好想想语言究竟是怎么一回事,
Chinese, Traditional	看看語言到底是怎麼一回事
Dutch	om te kijken wat taal eigenlijk is.
French	et examiner ce qu'est vraiment le langage.
German	und untersuchen, was Sprache wirklich ist.
Italian	e vedere che tipo di linguaggio è veramente,
Japanese	言語の本質を見る必要があります
Korean	언어란 과연 무엇인지에 대해 생각해 볼 필요가 있습니다.
Portuguese, Brazilian	e analisarmos o que a língua realmente é,
Romanian	şi să înţelegem ce e de fapt limbajul.
Russian	нужно понять, что такое есть сам язык,
Spanish	y observar qué es en realidad el idioma.

List of translation languages

15 languages

(languages written L-to-R to which more than 1,500 talks have been translated)

Bulgarian	1,567 talks	Korean	1,697 talks
Chinese, Simplified	1,697 talks	Portuguese	1,026 talks
Chinese, Traditional	1,634 talks	Portuguese, Brazilian	1,593 talks
Dutch	1,448 talks	Romanian	1,630 talks
French	1,572 talks	Russian	1,744 talks
German	1,527 talks	Spanish	1,667 talks
Italian	1,674 talks	Turkish	1,495 talks
Japanese	1,600 talks		

Note: some translations released by TED are not imported to TCSE for technical reasons

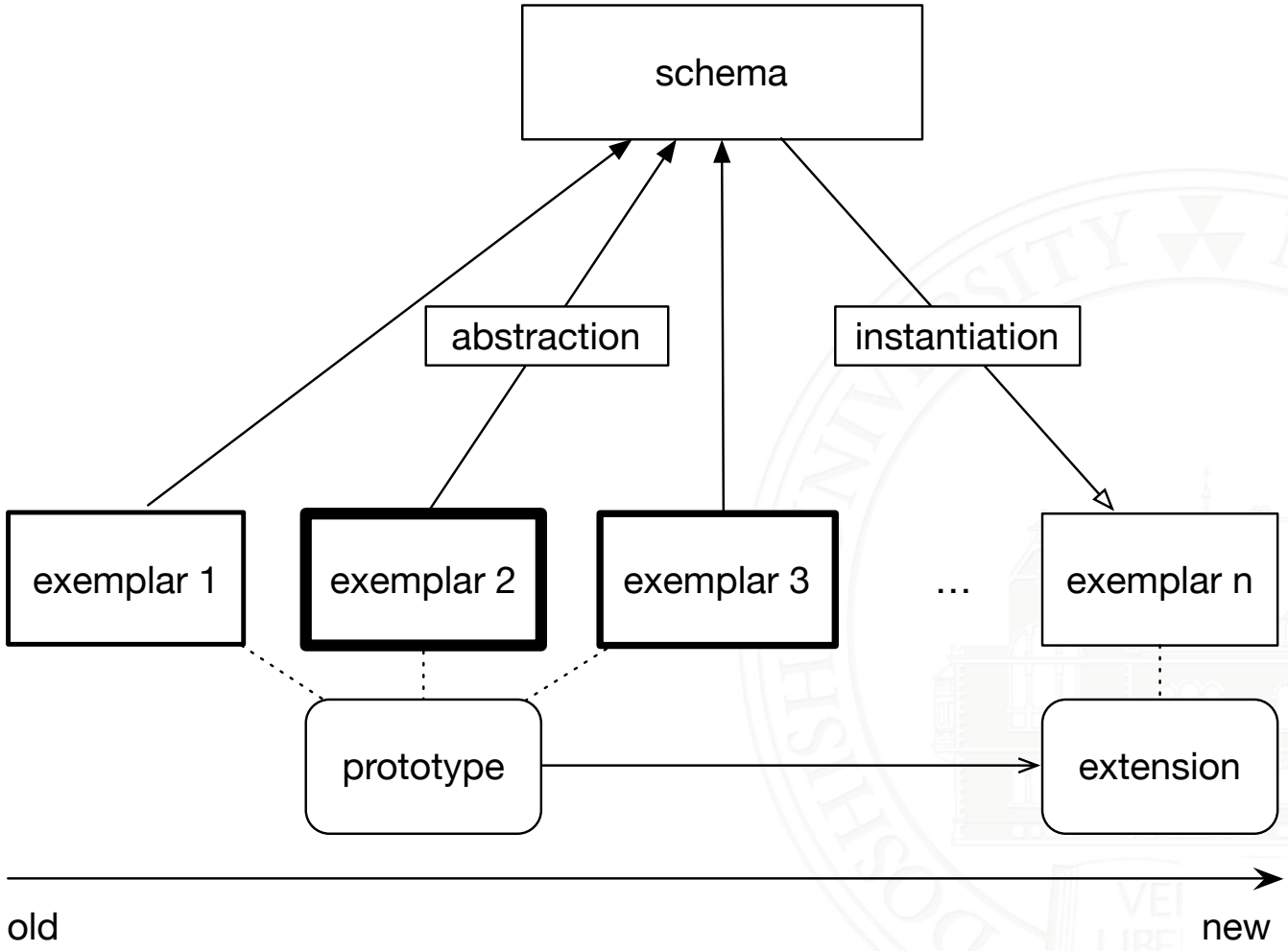
Usage-based model of language

Basic theoretical concept behind TCSE

→ usage-based model of language in terms of cognitive linguistics (cf. Langacker 1987, 1991, 2008; Barlow and Kemmer 2000; McEnery and Hardie 2011)

> The usage-based thesis holds that the mental grammar of the language user … is formed by the abstraction of symbolic units from situated instances of language use: an utterance. (Evans 2007: 216-217)

Schemas and instances/exemplars



Getting (good) exemplars from corpus

For linguistic research

TCSE provides "situated" instances of expressions

→ especially important in:

- cognitive linguistics
- discourse analysis

For language learning/teaching

TCSE offers readily available real samples of:

- usages of words, phrases, constructions, etc.
- possible translations of particular expressions

Conclusion

TED Corpus Search Engine (TCSE)

- available online at http://yohasebe.com/tcse
- searches more than 1,800 TED talk transcripts in English and translations in 15 languages
- designed and implemented based on usage-based model of language

Thank you!

Yoichiro Hasebe (Doshisha University, Kyoto, Japan)
yohasebe@gmail.com